Of arguments and wagers

Paul Christiano Follow
Dec 14, 2014 · 7 min read

(In which I explore an unusual way of combining the two.)

Suppose that Alice and Bob disagree, and both care about Judy's opinion. Perhaps Alice wants to convince Judy that raising the minimum wage is a cost-effective way to fight poverty, and Bob wants to convince Judy that it isn't.

If Judy has the same background knowledge as Alice and Bob, and is willing to spend as much time thinking about the issue as they have, then she can hear all of their arguments and decide for herself whom she believes.

But in many cases Judy will have much less time than Alice or Bob, and is missing a lot of relevant background knowledge. Often Judy can't even understand the key considerations in the argument; how can she hope to arbitrate it?

Wagers

For a warm-up, imagine that Judy could evaluate the arguments if she spent a long enough thinking about them.

To save time, she could make Alice and Bob wager on the result. If both of them believe they'll win the argument, then they should be happy to agree to the deal: "If I win the argument I get $100; if I lose I pay $100." (Note: by the end of the post, no dollars will need to be involved.)

If either side isn't willing to take the bet, then Judy could declare the case settled without wasting her time. If they are both willing to bet, then Judy can hear them out and decide who she agrees with. That person "wins" the argument, and the bet: **Alice and Bob are betting about what Judy will believe, not about the facts on the ground**.

Of course we don't have to stick with 1:1 bets. Judy wants to know the probability that she will be convinced, and so wants to know at what odds the two parties are both willing to bet. Based on that probability, she can decide if she wants to hear the arguments.

It may be that both parties are happy to take 2:1 bets, i.e. each believes they have a 2/3 chance of being right. What should Judy believe? (In fact this should always happen at small stakes: both participants are willing to pay some premium to try to convince Judy. For example, no matter what Alice believes, she would probably be willing to take a bet of $0.10 against $0.01, if doing so would help her convince Judy.)

If this happens, there is an arbitrage opportunity: Judy can make 2:1 bets with both of them, and end up with a guaranteed profit. So we can continuously raise the required stakes for each wager, until either (1) the market approximately clears, i.e. the two are willing to bet at nearly the same odds, or (2) the arbitrage gap is large enough to compensate Judy for the time of hearing the argument. If (2) happens, then Judy implements the arbitrage and hears the arguments. (In this case Judy gets paid for her time, but the pay is independent of what she decides.)

Recursion

Betting about the whole claim saved us some time (at best). Betting about parts of the claim might get us much further.

In the course of arguing, Alice and Bob will probably rely on intermediate claims or summaries of particular evidence. For example, Alice might provide a short report describing what we should infer from study Z, or Bob might claim “The analysis in study Z is so problematic that we should ignore it.”

Let’s allow anyone to make a claim at any time. But if Alice makes a claim, Bob can make a counterclaim that he feels better represents the evidence. Then we have a recursive argument to decide which version better represents the evidence.

The key idea is that **this recursive argument can also be settled by betting**. So one of two things happens: (1) Judy is told the market-clearing odds, and can use that information to help settle the original argument, or (2) there is an arbitrage opportunity, so Judy hears out the argument and collects the profits to compensate her for the time.

This recursive argument is made in context: that is, Judy evaluates which of the two claims she feels would be a more helpful summary within the original argument. Sometimes this will be a question of fact about which Alice and Bob disagree, but sometimes it will be a more complicated judgment call. For example, we could even have a recursive argument about which wording better reflects the nuances of the situation.

When making this evaluation, Judy uses facts she learned over the course of the argument, but she interprets the claim as she would have interpreted it at the beginning of the argument. For example, if Bob asserts "The ellipsoid algorithm is efficient" and Alice disagrees, Bob cannot win the argument by explaining that "efficient" is a technical term which in context means "polynomial time"—unless that's how Judy would have understood the statement to start with.

This allows Judy to arbitrate disagreements that are too complex for her to evaluate in their entirety, by showing her what she "would have believed" about a number of intermediate claims, if she had bothered to check. Each of these intermediate claims might itself be too complicated for Judy to evaluate directly—if Judy needed to evaluate it, she would use the same trick again.

Betting with attention

If Alice and Bob are betting about many claims over the course of a long argument, we can replace dollars by "attention points," which represent Judy's time thinking about the argument (perhaps 1 attention point = 1 minute of Judy's time). Judy considers an arbitrage opportunity "good enough" if the profit is more than the time required to evaluate the argument. The initial allocation of attention points reflects the total amount of time Judy is willing to spend thinking about the issue. If someone runs out of attention points, then they can no longer make any claims or use up any of Judy's time.

This removes some of the problems of using dollars, and introduces a new set of problems. The modified system works best when the total stock of attention points is large compared to the number at stake for each claim. Intuitively, if there are N comparable claims to wager about, the stakes of each should not be more than a 1/sqrt(N) of the total attention pool—or else random chance will be too large a factor. This requirement still allows a large gap between the time actually required to evaluate an argument (i.e. the initial bankroll of attention points) and the total time that would have been required to evaluate all of the claims made in the argument (the total stake of all of the bets). If each claim is itself supported by a recursive argument, this gap can grow exponentially.

Talking it out

If Alice and Bob disagree about a claim (rather, if they disagree about Judy's probability of accepting the claim) then they can have an incentive to "talk it out" rather than bringing the dispute to Judy.

For example, suppose that Alice and Bob each think they have a 60% chance of winning an argument. If they bring in Judy to arbitrate, both of them will get unfavorable odds. Because the surplus from the disagreement is going to Judy, both parties would be happy enough to see their counterparty wise up (and of course both would be happy to wise up themselves). This creates room for positive sum trades.

Rather than bringing in Judy to arbitrate their disagreement, they could do further research, consult an expert, pay Judy attention points to hear her opinion on a key issue, talk to Judy's friends—whatever is the most cost-effective way to resolve the disagreement. Once they have this information, their betting odds can reflect it.

An example

Suppose that Alice and Bob are arguing about how many trees are in North America; both are experts on the topic, but Judy knows nothing about it.

The easiest case is if Alice and Bob know all of the relevant facts, but one of them wants to mislead Judy. In this case, the truth will quickly prevail. Alice and Bob can begin by breaking down the issue into "How many trees are in each of Canada, the US, and Mexico?" If Alice or Bob lie about any of these estimates, they will quickly be corrected. Neither should be willing to bet much for a lie, but if they do, the same thing will happen recursively—the question will be broken down into "how many trees are east and west of the Mississippi?" and so on, until they disagree about how many trees are on a particular hill—a straightforward disagreement to resolve.

In reality, Alice and Bob will have different information about each of these estimates (and geography probably won't be the easiest way to break things down—instead they might combine the different considerations that inform their views, the best guess suggested by different methodologies, approximate counts of each type of tree on each type of land, and so on). If Alice and Bob can reach a rational consensus on a given estimate, then Judy can use that consensus to inform her own view. If Alice and Bob can't resolve their disagreement, then we're back to the previous case. The only difference is that now Alice and Bob have probabilistic disagreements: if Alice disagrees with Bob she doesn't expect to win the ensuing argument with 100% probability, merely with a high probability.

Odds and ends

This writeup leaves many details underspecified. In particular, how does Judy estimate how long it will take her to arbitrate a disagreement? This can be handled in several ways: by having Judy guess, by having Alice and Bob bet on the length of time until Judy reaches a conclusion, by having them make bets of the form "Alice will agree with me with Z effort," or so on. I don't know what would work best.

Despite my use of the word "recursion," the estimate for "time to settle an argument" (which Judy uses to decide when the stakes are high enough to step in and resolve a disagreement) probably shouldn't include the time required to settle sub-arguments, since Judy is being paid separately for arbitrating each of those. The structure of the arguments and sub-arguments need not be a tree.

This is a simple enough proposal that it can be realistically implemented, so eventually we'll hopefully see how it works and why it fails.

I expect this will work best if Alice and Bob often argue about similar topics.

This scheme was motivated by a particular exotic application: delegating decision-making to very intelligent machines. In that setting the goal is to scale to very complex disagreements, with very intelligent arguers, while being very efficient with the overseer's time (and more cavalier with the arguers' time).