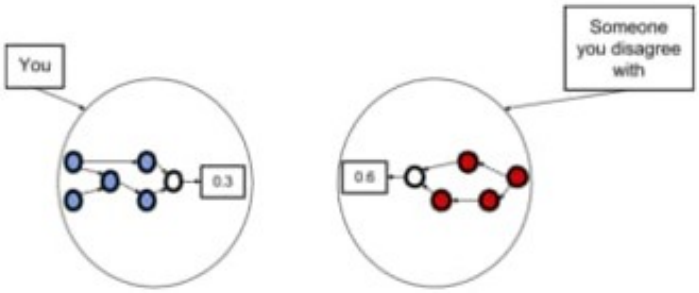

You
0.3
Someone you disagree with
0.6

Share Models, Not Beliefs

A Sketch of Good Communication

"Often I compare my own Fermi estimates with those of other people, and that's sort of cool, but what's way more interesting is when they share what variables and models they used to get to the estimate."

– Oliver Habryka*, at a model building workshop at FHI in 2016*

One question that people in the AI x-risk community often ask is

> "By what year do you assign a 50% probability of human-level AGI?"

We go back and forth with statements like "Well, I think you're not updating enough on AlphaGo Zero." "But did you know that person X has 50% in 30 years? You should weigh that heavily in your calculations."

However, 'timelines' is not the interesting question. The interesting parts are in the causal models *behind* the estimates. Some possibilities:

- Do you have a story about how the brain in fact implements back-propagation, and thus whether current ML techniques have all the key insights?
- Do you have a story about the reference class of human brains and monkey brains and evolution, that gives a forecast for how hard intelligence is and as such whether it's achievable this century?
- Do you have a story about the amount of resources flowing into the problem, that uses factors like 'Number of PhDs in ML handed out each year' and 'Amount of GPU available to the average PhD'?

Timelines is an area where many people discuss one variable all the time, where in fact the interesting disagreement is much deeper. Regardless of whether our 50% dates are close, when you and I have different models we will often recommend contradictory strategies for reducing x-risk.

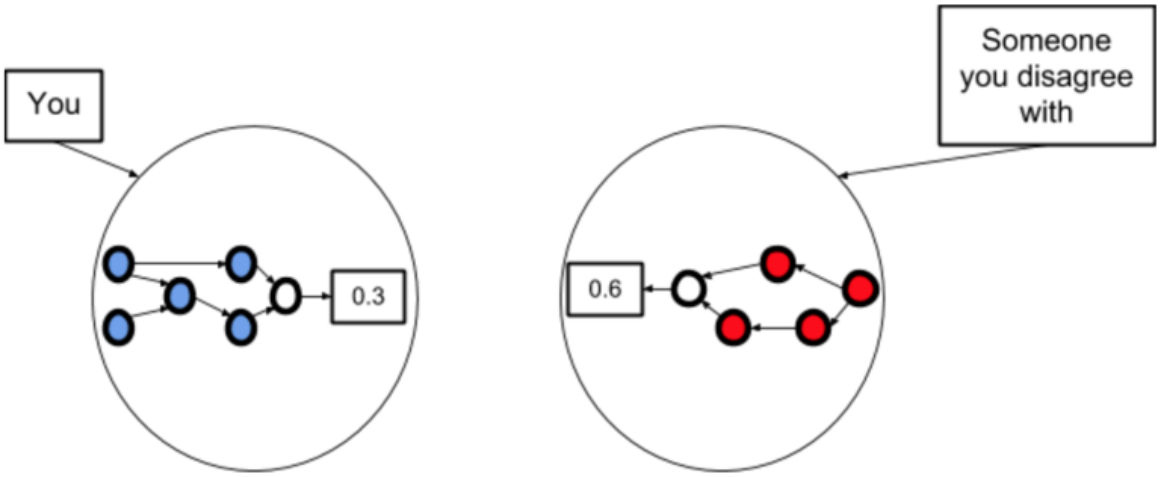
For example, Eliezer Yudkowsky, Robin Hanson, and Nick Bostrom all have different timelines, but their models tell such different stories about what's happening in the world that focusing on timelines instead of the broad differences in their overall pictures is a red herring.

(If in fact two very different models converge *in many places*, this is indeed evidence of them both capturing the same thing - and the more different the two models are, the more likely this factor is 'truth'. But if two models significantly disagree on strategy and outcome yet hit the same 50% confidence date, and we should not count this as agreement.)

Let me sketch a general model of communication.

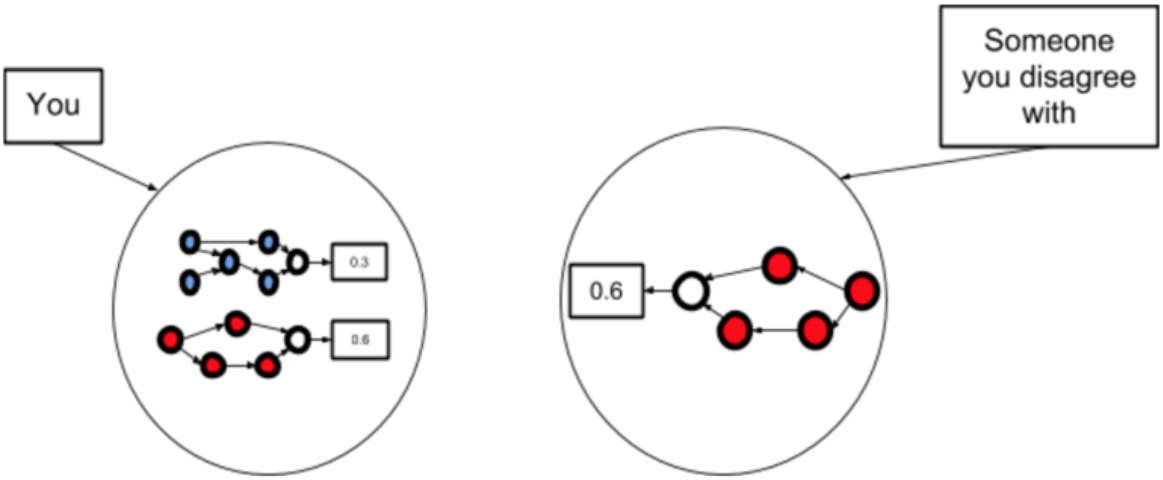
A Sketch

Step 1: You each have a different model that predicts a different probability for a certain event.



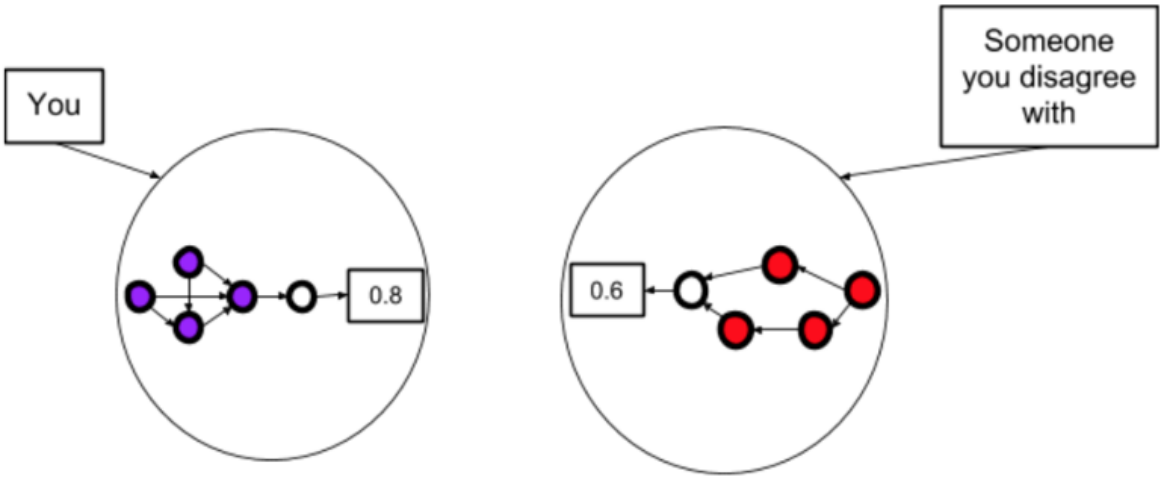
"I see your probability of reaching human level AGI in the next 25 years is 0.6, whereas mine is 0.3."

Step 2: You talk until you have understood how they see the world.



"I understand that you think that all the funding and excitement means that the very best researchers of the next generation will be working on this problem in 10 years or so, and you think there's a big difference between a lot of average researchers versus having a few peak researchers."

Step 3: You do some cognitive work to integrate the evidences and ontologies of you and them, and this implies a new probability.



> “I have some models from neuroscience that suggest the problem is very hard. I'd thought you thought the problem was easy. But I agree that the greatest researchers (Feynmans, von Neumans, etc) can make significantly bigger jumps than the median researcher."
>
> "If we were simply increasing the absolute number of average researchers in the field, then I’d still expect AGI much slower than you, but if now we factor in the very peak researchers having big jumps of insight (for the rest of the field to capitalise on), then I think I actually have shorter timelines than you.”

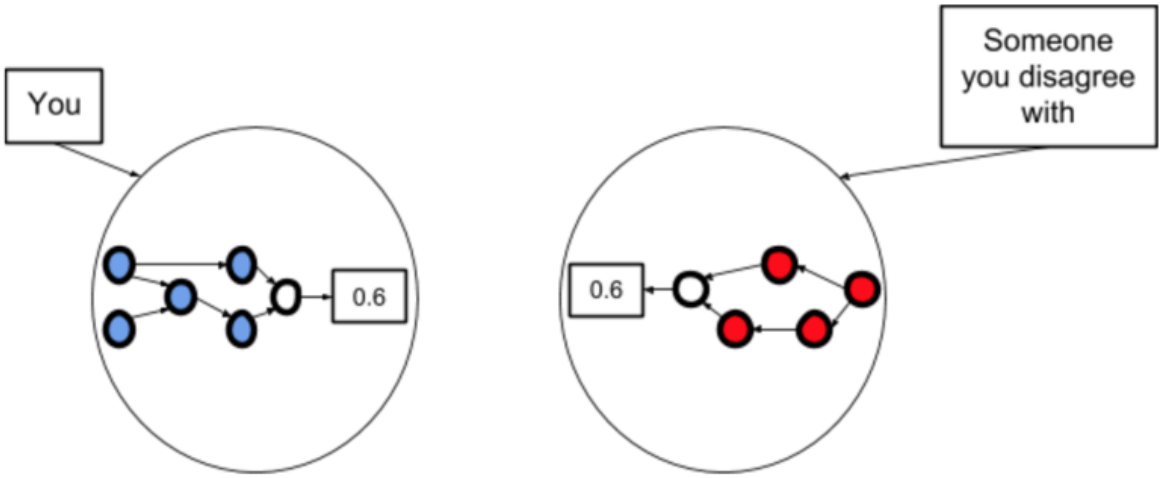
One of the common issues I see with disagreements in general is people jumping prematurely to the third diagram before spending time getting to the second one. It’s as though if you both agree on the decision node, then you must surely agree on all the other nodes.

I prefer to spend an hour or two sharing models, before trying to *change* either of our minds. It otherwise creates false consensus, rather than successful communication. Going directly to Step 3 can be the right call when you’re on a logistics team and need to make a decision quickly, but is quite inappropriate for research, and in my experience the most important communication challenges are around deep intuitions.

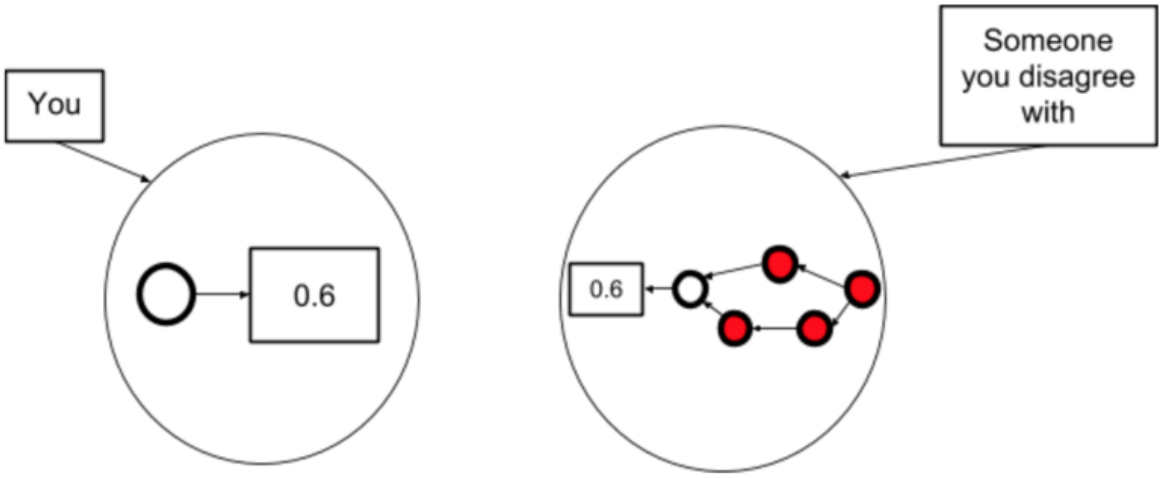
Don't practice coming to agreement; practice exchanging models.

Something other than Good Reasoning

Here’s an alternative thing you might do after Step 1. This is where you haven't changed your model, but decide to agree with the other person anyway.



This doesn't make *any* sense but people try it anyway, especially when they're talking to high status people and/or experts. "Oh, okay, I'll try hard to believe what the expert said, so I look like I know what I'm talking about."



This last one is the *worst*, because it means you can't notice your confusion any more. It represents "Ah, I notice that p = 0.6 is inconsistent with my model, therefore I will *throw out my model*." Equivalently, "Oh, I don't understand something, so *I'll stop trying.*"

This is the first post in a series of short thoughts on epistemic rationality, integrity, and curiosity. My thanks to Jacob Lagerros and Alex Zhu for comments on drafts.

Descriptions of your experiences of successful communication about subtle intuitions (in any domain) are welcomed.

Hold On To The Curiosity

I.

Recently, an excited friend was telling me the story behind why we care about the mean, median and mode.

They explained that a straightforward idea for what you might want in an ‘average’ number, is something that minimises how far it is from all the other numbers in the dataset - so if your numbers are 1, 2 and 3, you want a number x such that the sum of the distance to each datapoint is as small as possible. It turns out this number is 2.

However, if your numbers are 1, 2, and 4, the number that minimises the distance from all of them is *also* 2.

Huh?

When my friend told me this, the two other people I was with sort of said “Okay”. I said “What? No! I don’t believe you! It has to change when the data does - it’s a linear sum, so it has to change! It’s like you’re saying the sum of 1, 2 and 3 is the same as the sum of 1, 2 and 4. This is just *wrong*." Suffice to say, my friend’s claim wasn’t predicted by my understanding of math.

Now, did I really not believe my friend? The other two people with us were certainly fine with it. Isn’t this just *bayesianism*? That’s how the old joke goes:

> **Math teacher:** Now I’m going to prove to you that X is true.
>
> **Bayesian:** You just did.

Actually, no. You taught me a detail to memorise, but my models didn’t improve. I won’t be able to improve how I use averages, because I don’t understand how it fits in with everything else I understand - it doesn’t fit with the models I use everywhere else in math.

I mean, I could’ve nodded along. It’s only one fact, after all. But if I’m going to remember it in the long term, it should connect to my other models and be reinforced. The alternative is to be stored in the brain with all those other memorised facts that students learn for exams and forget immediately after.

If you’re trying to build new models of a domain, it’s important to choose to speak from the confusion, not from the rest of yourself. Don’t have conversations about whether you believe a thing. Instead talk about whether you understand it.

(The problem above was the definition of the median, and an explanation of the math for the curious can be found in this comment.)

II.

It can be really hard to *feel* your models. Qiaochu Yuan’s method of learning involves ramping feeling-his-models up to 11. I recall him telling me about trying to learn what fire was once, where his first step was to just really feel his confusion:

> What the hell is this orange stuff? How on earth does it get here? Why is it flickering? WHAT IS FIRE?!

After feeling the confusion, Qiaochu holds onto his *frustration* (which he finds easier to hold), and tries throwing ideas and possible explanations at it until all the parts finally fit together - that feeling when you say "Ohhhhhhh" and the models finally compute, and your beliefs predict the experience you have. *Be frustrated with reality*.

Tim Urban (of WaitButWhy) tells a similar story, where he can only write essays about things he *doesn't currently understand* - and as he's digging through all the facts and pieces things together, he writes down the things that made sense to him, that would successful get the models across to an earlier version of Tim Urban.

I used to think this made no sense and he must just be bad at introspecting - shouldn't you have to build an excellent model of other people to write so compellingly for so many tens of thousands of them?

Yet it's actually really rare for authors to be strongly connected to *their own* models - when a teacher explains something for the hundredth time, they likely can't remember what it was like to learn it for the first. And so Tim's explanations can be clearer than most.

In the opening example where I was surprised by the definition of the median, if you had offered me a bet I would've bet on the side that this was the definition of a median. But it was not a useful thought for me in that moment, to set aside my confusion and say "On reflection I believe you". It can be correct in conversation, when your goal is understanding, to hold onto the confusion, the frustration, and let your models do the speaking.

III.

I often feel people try to move a conversation toward whether I believe the claim, rather than discussing and sharing what we each understand.

> "Do you *believe m*e when I say picking an average by minimising the distance to all the points is the same as the median?
>
> "Hmm, can you tell me *why* that's the case? I have a model of arithmetic that says it shouldn't be…"

A phrase I often use: *"You may have changed my betting odds but you haven't changed my models!*"

We're all in the game of trying to build models. Whether you're trying to understand the field of science you're attempting to add knowledge to, the product your startup is building, or the architecture of the AGI you're trying to align, you need good models to leverage reality for whatever you care about.

One of the most important skills in life is the ability to hold onto your confusion and let your models do the talking, so they can interface with reality more directly. Choosing to notice and hold on to your confusion is hard, and it's so easy to lose sight of it.

To put it another way, here are some perfectly acceptable noises to make when your goal is understanding:

What? No! I don’t believe you! That *can't* be true!

I expect that some but not all of this post is surprisingly Ben-specific. My thanks to Alex Zhu (zhukeepa) and Jacob Lagerros (jacobjacob) for reading drafts.

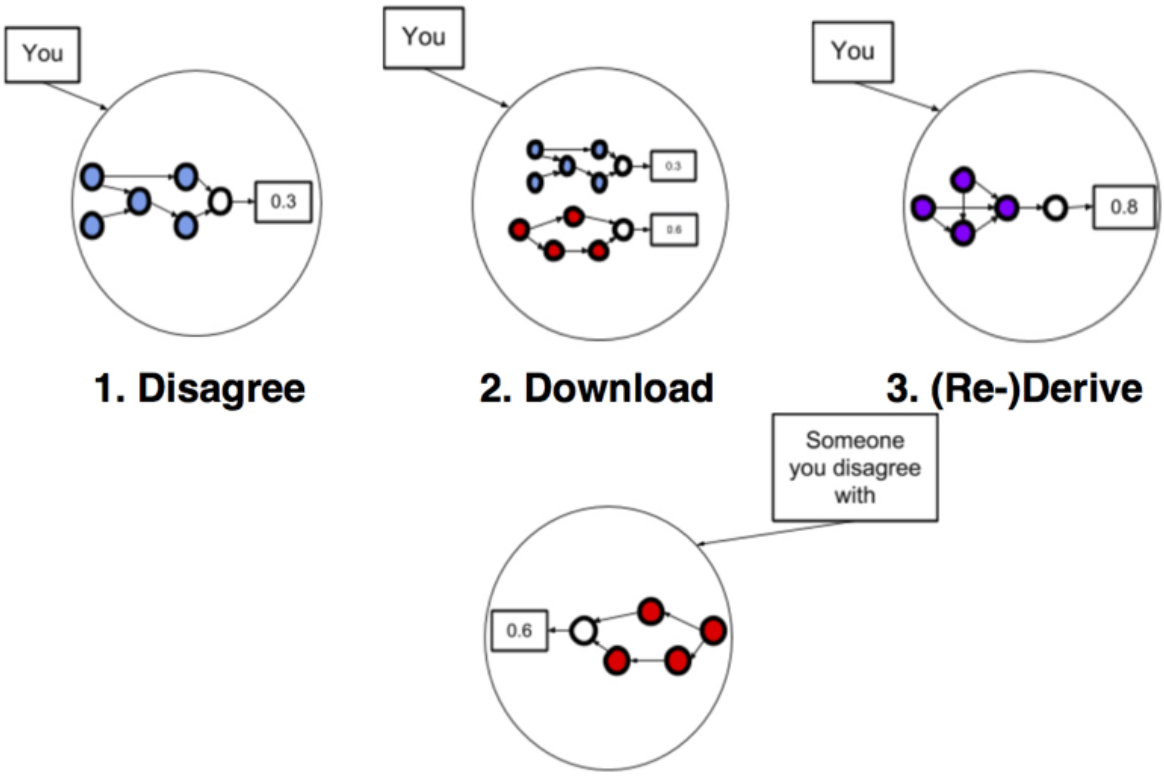
Form Your Own Opinions

Follow-up to: *A Sketch of Good Communication*

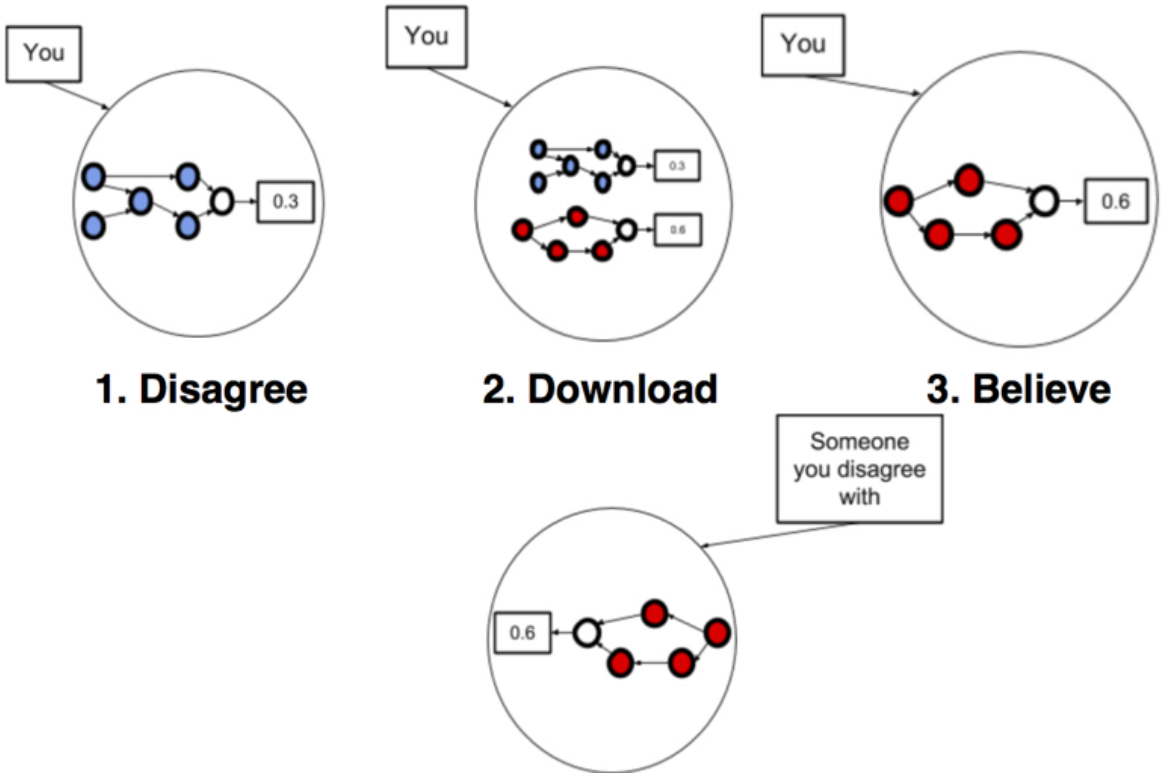
Question:

> Why should you integrate an expert's model with your model at all? Haven't you heard that people weigh their own ideas too heavily - you should just defer to them.

Here's a quick reminder of the three step process:



And this new proposal, I think, suggests changing what you do after step 2.



The people who have the best ideas, as it seems to me, often change their plans as a result of their debates in all the other fields. Here's three examples[1] of people doing this with AI timelines.

- **[Person 1]** Oh, AI timelines? Well, I recall reading that it took evolution 10^9 years to go from eukaryotes to human brains. I'd guess that human developers are about 10^6 times more efficient than evolution, so I expect it to take 1000 years to get there from the point where we built computers. Which puts my date at 2956.
    - If I'm wrong I'll likely have to learn something new about developers competence relative to evolution, or about how humans get to do a type of intelligence evolution wasn't allowed to for some reason.
- **[Person 2]** Oh, AI timelines? Well, given my experience working on coding projects, it seems to me that projects take 50% extra time to run than you'd expect once you've got the theory down, so I'll take the date by which we should have enough hardware to build a human brain, estimate the coding work required for the necessary project, and add 1.5x time to it.
    - And if you change my mind on this, it will help make my models of project time more accurate, and change how I do my job.
- **[Person 3]** Oh, AI timelines? Well, given my basic knowledge of GDP growth rates I'd guess that being able to automate this percentage of the workforce would cause a doubling every X unit of time, which I expect for us (at current rates) only to be able to do after K years.
    - If you show me I'm wrong it'll either be because GDP is not as reliable as I think, or because I've made a mistake extrapolating the trend as it stands.

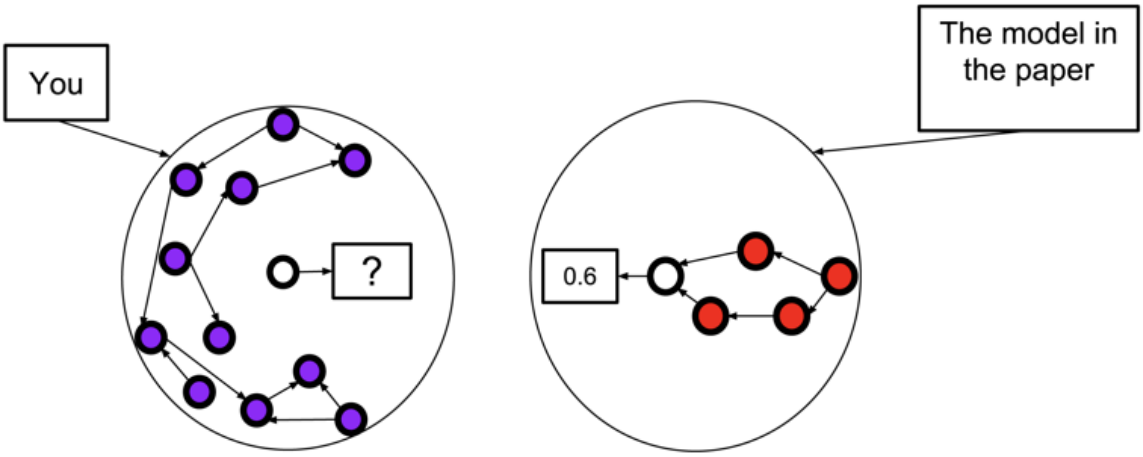
Can you see how a conversation between either of these two people would lead them both to learn not only about AI, but also about models of evolution/large scale coding projects/macroeconomics?

Recently, Jacob Lagerros and I were organising a paper-reading session on a recent Distill.pub paper, and Jacob was arguing for a highly-structured and detailed read-through of the paper. I wanted to focus more on understanding people's current confusions about the subfield and how this connected to the paper, rather than focusing solely on the details.

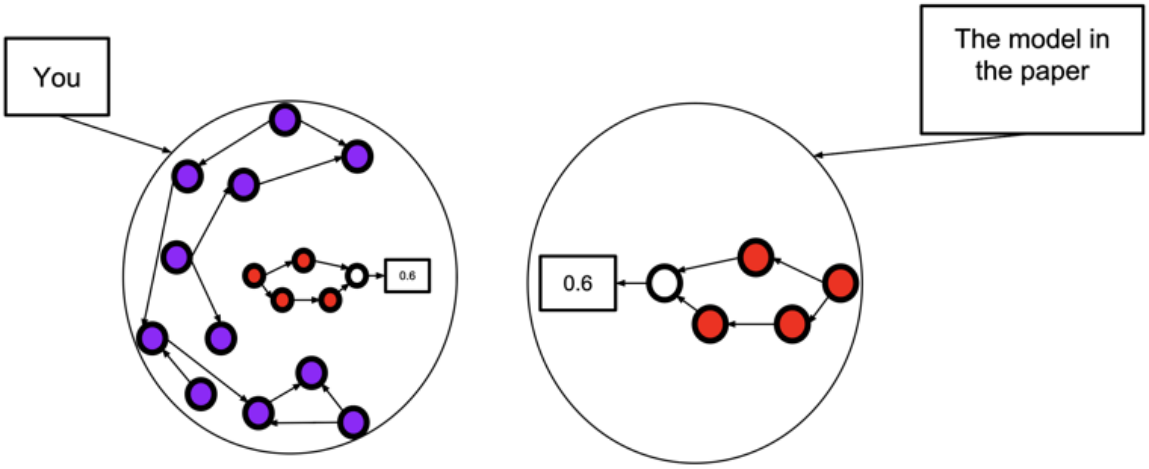
Jacob said "Sometimes though Ben, you just need to learn the details. When the AlphaGo paper came out, it's all well and good to try to resolve your general confusions about Machine Learning, but sometimes you just need to learn *how AlphaGo worked*."

I responded: "Quite to the contrary. When reading an *important* paper, this is an especially important time to ask high-level questions like 'is research direction X ever going to be fruitful' and 'is this a falsification of my current model of this subfield', because you rarely get evidence of that sort. We need people to load up their existing models, notice what they're confused about, and make predictions *first*."

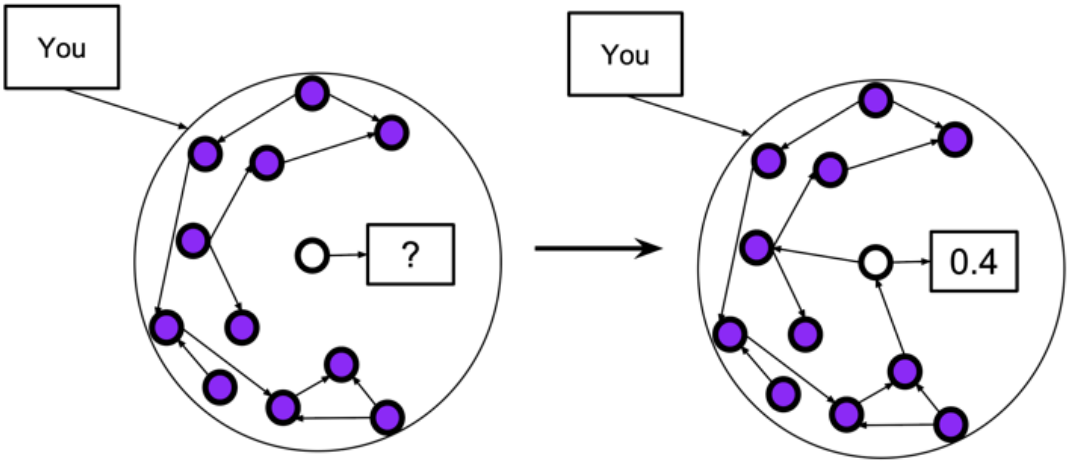
I wanted to draw diagrams with background models, where Jacob was arguing for:



Followed by:



But I was trying to say:



This is my best guess at what it *means* to have an opinion - to have an integrated causal model of the world that makes predictions about whatever phenomena you're discussing. It is basically steps −1 and 0 before the 3-step process outlined at the top.

The truth is somewhere in between, as it's not possible to have a model that connects to action without it in some way connecting to your other models. But it's always important to ask "What other parts of my models have implications here that this can give data on?"

Another way of saying this would be 'practice introspecting on your existing models and then building models of new domains'.

And I think a way to do this is first to form your own opinion, then download the other person's model, and finally integrate. Don't just download someone else's model first - empirically I find people have a real hard time imagining the world to be a different way once they've heard what you think, especially if you're likely to be right. Tell your friend you need a few minutes silence to think, build an opinion, then forge toward the truth together.

(One way to check if you're doing this in conversation, is to ask whether you're regularly repeating things you've said before, or whether you're often giving detailed explanations for the first time - whether you're thinking thoughts *you've not thought before*. The latter is a sign you're connecting models of diverse domains.)

I want to distinguish this idea from "Always give a snappy answer when someone asks you a question". It's often counterproductive to do that, as building models takes time and thought. But regularly do things like "Okay, let me think for a minute" *minute long silence* "So I predict that key variables here are..."

(Another way of saying that is 'Don't be a button-wall'. If you ever notice that someone is asking you a question and all you have are stock answers and explanations, try instead to think a new thought. It's much harder, but far more commonly leads to very interesting conversations where you actually learn something.)

Another way of saying 'form lots of opinions' is 'think for yourself'.

Summary

The opening question:

> Why is forming your own opinions better than simply downloading someone else's model (like an expert's)? Why should you even integrate it with your model at all?
>
> Surely what you want is to get to the truth - the right way of describing things. And the expert is likely closer to that than you are.
>
> If you could just *import* Einstein's insights about physics into your mind… surely you should do that?

Integrating new ideas with your current models is really valuable for several reasons:

- It lets your models of different domains *share* data (this can be really helpful for domains where data is scarce).
- It gives you more and faster feedback loops about new domains:
  - More interconnection → More predictions → More feedback
  - This does involve downloading the expert's model - it just adds extra sources of information to your understanding.
- It's the only way to further entangle yourself with reality. Memorising the words of the expert is not a thing that gives you feedback, and it's not something that's self-correcting if you got one bit wrong.

Footnotes

1. I've picked examples that require varying amounts of expertise. It's great to integrate your highly detailed and tested models, but you don't require the protection of 'expertise' in order to build a model that integrates with what you already know - I made these up and I know very little about how GDP or large scale coding projects work. As long as you're building and integrating, you're moving forwards.

2. I like to think of Tetlock's Fox/Hedgehog distinction through this lense. A fox is a person who tries to connect models from all different domains, and is happy if their model captures a significant chunk (e.g. 80%) of the variance in that domain. A hedgehog is someone who wants to download *the correct* model of a domain, and will refine it with details until it captures as much of the variance as possible.

Thanks to Jacob Lagerros (jacobjacob *) for comments on drafts.*

Goodhart Taxonomy: Agreement

> "When a measure becomes a target, it ceases to be a good measure."
>
> -Goodhart's Law

If you spend a while talking with someone you disagree with and end up agreeing, this is a sign you are both reasoning and communicating well - one of the primary uses of good reasoning is resolving disagreement. However, if you use agreement as your main proxy for good reasoning, some bad things might happen.

Scott Garrabrant has helpfully laid out four different models of how things can go wrong if you optimise for the proxy really hard, a phenomenon known as 'goodharting' (based on Goodhart's Law that any proxy when optimised for hard, stops being a good proxy). I want to take a look at each model and see what it predicts for the real world, in the domain of agreement.

Regressional

First, you can fall prey to **regressional goodharting**. This is when the proxy you're optimising for is a good measure of the thing you actually care about, but plus some noise (i.e. other uncorrelated variables), and the examples that maximise the sum of these are the examples that maximise the noise. I can think of three ways this could happen: misunderstanding, spurious correlation, and shared background models.

Misunderstanding is the simple idea that, of the times when you most agree with someone, you misunderstood each other (e.g. were using words differently). Especially if it's an important topic and most people disagree with you, suddenly the one person who gets you seems to be the best reasoner you know (if you're regressional goodharting).

Spurious correlation is like I described in A Sketch of Good Communication - two people who have different AI timelines can keep providing each other with new evidence until they have the same 50th percentile date, but it may turn out that they have wholly different causal models behind, and thus don't meaningfully agree around AI x-risk. This is different from misunderstanding, because you heard the person correctly when they stated their belief.

And shared background models happens like this: You decide that the people who are good reasoners and communicators are those who agree with you a lot on complex issues after disagreeing initially. Often this heuristic ends up finding people who are good at understanding your point of view, and updating when you make good arguments. But if you look at the people who agree with you the most of these people, you'll tend to start finding people who share a lot of background models. *"Oh, you've got a PhD in economics too? Well obviously you can see that these two elastic goods are on the pareto frontier if there's an efficient market. Exactly. We're so good at communicating!"*

Extremal

Second, you can fall prey to **extremal goodharting**. This is where the peaks of your heuristic are actually falling out of an entirely different process, and have no bearing at all on the thing you cared about. Here's some things you might actually get if you followed the heuristic 'agrees with me' to its extreme:

- A mirror

- A service sector worker whose rule is ‘the customer is always right’
- Someone who trusts you a lot personally and so believes what you say is true
- A partner who likes the sound of your voice and knows saying ‘yes, go on’ causes you to talk a lot
- An identical copy of you (e.g. an emulated mind)

While these don’t seem like practical mistakes any of us would make, I suppose it’s a good skill to be able to know the literal maximum of the function you wrote down. It can help you to not build the wrong AGI, for example.

Adversarial

But there is one particularly common type of process that can end up being spuriously high in your proxy: our third type, **adversarial goodharting**. This is where someone notices that you’ve connected your proxy to a decision over a large amount of resources, thus creating in them an incentive to disguise themselves as maximising your proxy.

You’ll often incentivise the people around you to find ways to agree with you more than finding ways to successfully communicate. If you have a person who you’ve not successfully communicated with who says so, and another who is in the same state but pretends otherwise, then you’ll prefer the liar.

People who are very flexible with their beliefs (i.e. don’t really have models) and good at sounding like they agree with you, will be rewarded the most. These are yes-men, they aren’t actually people who know how to update their beliefs on a fundamental level, and their qualities deeply do not correlate with the goal of ‘good communicator’ at all.

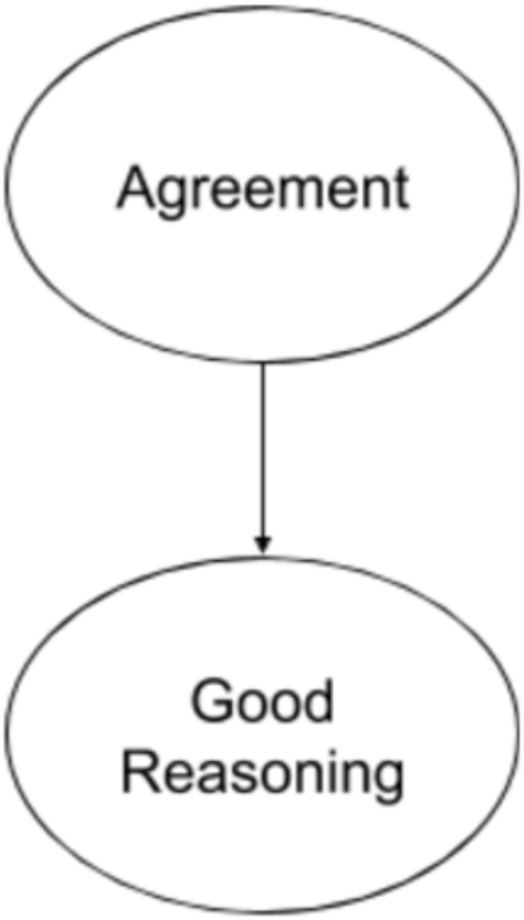
Adversarial goodharting can happen even if humans aren’t explicitly trying to be hostile. Sure, the liars looking for power will try to agree with you more, but a perfectly well-intentioned manager goodharting on agreement will try to get more power, entirely because they observe it leads to them agreeing with people more, and that this is a sign of good reasoning.

This is most clear if you are a manager. If you’re the boss, people have to agree with you more. If you’re using agreement as your main proxy for good communication and honestly not attempting to grab power, you’ll nonetheless learn the pattern that taking power *causes you to be a good communicator and reasoner*. And I don’t think this is at all unlikely. I can well imagine this happening by accident at the level of social moves. *“Huh, I notice when I stand up and speak forcefully, people agree with me more. This must be making me a better communicator - I’ll do this more!”*

Causal

Fourthly and finally, we come to most juicy type of goodharting in this domain, **causal goodharting**. This is where you have the causal model the wrong way around - you notice that basketball players are taller than other people, so you start playing basketball in an effort to get taller.

If you causal goodhart on agreement, you don’t believe that good reasoning and communication cause agreement, but the opposite. You believe that agreement causes good reasoning. And so *you try to directly increase the amount of agreement*in an attempt to reason better.



Modelling people as having this belief predicts many experiences I've had in the real world.

It seems to me that causal goodharting is best understood by the beliefs it leads to. Here are three, followed by some bulleted examples of what the beliefs can lead to.

The first belief is **if a reasoning process doesn't lead to agreement, that's a bad process**.

- You'll consider an extended session of discussion (e.g. two hours, two days) to be a failure if you don't agree at the end, a success if you do, and not measure things like "*I learned a bunch more about how Tom thinks about management"* as positive hits or "*It turned out we'd been making a basic mistake for hours*" as negative hits.

The second belief is **if I disagree with someone, I'm bad at reasoning**.

- If someone has expressed uncertainty about a question, I'll be hesitant to express a confident opinion, because then we'll not be in agreement and that means I'm a bad communicator.

- If it happens the other way around, and you express a confident opinion after I've expressed uncertainty, I'll feel an impulse to say *"Well that's certainly a reasonable opinion"* (as opposed to *"That seems like the wrong probability to have"*) because then it sounds like we agree at least a bit. In general, when causal goodharting, people will feel uncomfortable having opinions - if you disagree with someone it's a signal you are a bad communicator.
- You'll only have opinions either when you think the trade-off is worth it (*"I see that I might look silly, but no, I actually care that we check the exhaust is not about to catch fire"*) or when you have a social standing such that people will defer to you (*"Actually, if you are an expert, then your opinion in this domain gets to be right and we will agree with you"*) - that way you can be free from the worry of signalling you're bad at communication and coordination.

In my own life I've found that treating someone as an 'expert' - whether it's someone treating me, or me treating someone else - lets that person express their opinions more and without obfuscation or fear. It's a social move that helps people have opinions *"Please meet my friend Jeff, who has a PhD in / has thought very carefully about X."* If I can signal that I defer to Jeff on this topic, then the social atmosphere can make similar moves, and stop Jeff being afraid and *actually think*.

(My friend notes that sometimes this goes the other way around, that sometimes people are much more cautious when they feel they're being tested on knowledge they're supposed to be an expert on. This is true, I was talking about the opposite, which happens much more when the person is around a group of laymen without their expertise.)

The third belief is **if someone isn't trying to agree, they're bad at reasoning**.

- I'm often in situations where, if at the end of the conversation people don't say things like *"Well you made good points, I'll have to think about them"* or *"I suppose I'll update toward your position then"* they're called 'bad at communicating'.

(An umeshism I have used many times in the past: If you never ending hangouts with friends saying *"Well you explained your view many times and I asked clarifying questions but I still don't understand your perspective"* then you're not trying hard enough to understand the world - and instead are caring too much about signalling agreement.)

- If you disagree with someone who signals that they are confident in their belief and are unlikely to change their mind, you'll consider this a sign of being bad at communication, even if they send other signals of having a good model of why you believe what you believe. Basically, people who are right and confident that they're right, can end up looking like bad reasoners. The normal word I see used for these people is 'overconfident'.

(It's really weird to me how often people judge others as overconfident after having one disagreement with them. Overconfidence is surely something you can only tell about a person after observing 10s of judgements.)

- Once you successfully communicate (*"Oh, I understand your perspective now")* you'll expect you also have to agree, rather than have a new perspective that's different from your or their prior beliefs. *"Well, I guess I must agree with you now..."*

A general counterargument to many of these points, is that all of these *are genuine signs of bad reasoning or bad communication*. They are more likely to be seen in world where you or I are a bad reasoner than if we're not, so they are bayesian evidence. But the problem I'm pointing to is that, if your model only uses this heuristic, or if it takes it as the most important heuristic that accounts for 99% of the variance, then it will fail *hard* on these edge cases.

To take the last example in the scattershot list of causal goodharting, you might assign someone who is reasoning perfectly correctly, as overconfident. To jump back to regressional goodharting, you'll put someone who just has the same background models as you *at the top of your list of good reasoners*.

Overall, I have many worries about the prospect of using agreement as a proxy for good reasoning. I'm not sure of the exact remedy, though one rule I often follow is: Respect people by giving them your time, not your deferrence. If I think someone is a good reasoner, I will spend hours or days trying to understand the gears of their model, and disagreeing with them fiercely if we're in person. Then at the end, after learning as much as I can, I'll use whatever moving-parts model I eventually understand, using all the evidence I've learned from them and from other sources. But I won't just repeat something because someone said it.

My thanks to Mahmoud Ghanem for reading drafts.