

What Would You Build Instead?

Almost all the argument about AI is about whether the current approach will reach artificial general intelligence. Very little of it is about whether AGI is the right thing to be aiming at. This is the second question, and it has a better answer.



WHERE THIS SITS

There is more language in the world than anyone can attend to. The obvious fix, an AI that reads it all for you and tells you what is true, does not work: the judgement that would take requires having been wrong at your own expense, and nothing a model does costs it anything. So the judging stays with people, and almost nothing we currently inspect can tell whether a person actually judged or simply agreed.

01 Why Does It All Feel Like Too Much?

02 Why Can't You Just Be Told?

03 What Holds a Claim Up?

04 How Would You Know?

05 Who Pays If It's Wrong?

06 **What Would You Build Instead?**

Six short pieces, each drawn from a working paper. Each one stands on its own; this is the sixth. They are ordered so that every piece ends by raising the question the next one is called.

The question nobody is asking

Will machines reach human-level general intelligence, and when? That question has hundreds of billions of dollars behind it and most of the argument.

Here is the other one. Suppose it works. Is that the thing we were trying to build?

THIS PIECE ASSUMES THE DIAGNOSIS

A companion piece, *Why Can't You Just Be Told?*, sets out why a system trained on text cannot acquire the judgement a practitioner acquires by being wrong. That argument is not repeated here. This one takes it as given and asks what follows for anybody actually building something.

Three things wrong with the goal

THE BENCHMARK PROBLEM

We measure progress with tests, and the tests keep being passed by systems nobody would trust to do the actual job. The test was meant to stand in for the ability. Passing it turns out not to be the same as having it.

THE RETURNS PROBLEM

Each increment costs more and buys less. That is an observation about the curve, not a prediction about the ceiling.

THE DISPLACEMENT PROBLEM

The goal has quietly moved. We started out asking how to help people think better, and ended up asking how to build something that does not need them.

The first two are arguments about the method. The third is an argument about the destination, and it is the one that matters.

How the goal moved

The field did not start here.

LICKLIDER, 1960

Wrote about man-computer symbiosis: a partnership in which each side did what it was good at.

ENGELBART, 1962

Proposed augmenting human intellect. Not replacing the person, but raising what a person could do.

Somewhere between then and now the aim became autonomy. Not a better partner, but a thing that does not need a partner.

Nobody voted on that. It happened because autonomy is easier to measure, easier to sell, and easier to describe in a funding round than a partnership is.

Replace, or amplify

This is the whole constructive argument in one distinction, and once you see it you cannot unsee it.

TO REPLACE A FUNCTION

You have to reproduce that function. All of it. Every part, to a standard at least as good, or the replacement is worse than what it replaced.

TO AMPLIFY A PERSON

You do not have to reproduce anything. You find the one place the person is actually stuck, and widen that. Everything else is left alone, because the person is still there doing it.

The first is five unsolved problems at once. The second is one problem, and it is the one that was slow anyway.

The same job, two designs

The left column is why the hard version is hard. The right column is why the useful version is tractable this year.

TO REPLACE THE JOB

gather what is known	BUILD IT
spot what is missing	BUILD IT
weigh it against the case	BUILD IT
decide	BUILD IT
answer for the decision	BUILD IT

five problems, all unsolved

TO AMPLIFY THE PERSON

gather what is known	
spot what is missing	THE PERSON
weigh it against the case	THE PERSON
decide	THE PERSON
answer for the decision	THE PERSON

one problem, the one that was slow

Glasses do not see for you

Every tool that has genuinely extended what people can do works this way, and none of them works the other way.

GLASSES

Do not see on your behalf. They fix one specific failure in one specific part of seeing, and leave the rest of your eye alone.

A CALCULATOR

Does not do your mathematics. It removes the arithmetic so you can spend the attention on which sum to do and whether the answer is sensible.

A WHEEL

Does not travel for you. It changes the cost of a step so that a journey that was impossible becomes ordinary.

Not one of these had to reproduce the thing it helped. That is not a coincidence, it is the reason they worked.

So what is the bottleneck?

If the design question is which single thing to widen, then it matters a great deal what is actually narrow. And the thing in short supply is not intelligence.

It is attention. There is more worth reading than anybody can read, more worth checking than anybody can check, and no amount of raw capability on the machine's side adds a single hour to the person's day.

A system designed to produce more output is widening something that was never the constraint. A system designed to make a person's attention go further is widening the thing that was.

Five things such a system would do

01 Built around the work, not the model.

The starting question is what the person is trying to do, not what the system happens to be able to produce.

02 Knows the situation, not just the question.

What is at stake, how reversible this is, who carries it if it is wrong. A system that cannot tell a draft from a decision should treat everything as a decision.

03 The decision stays with the person, by design.

Not as a policy that can be turned off under deadline pressure. As a property of how the thing is built.

04 Protects what builds judgement.

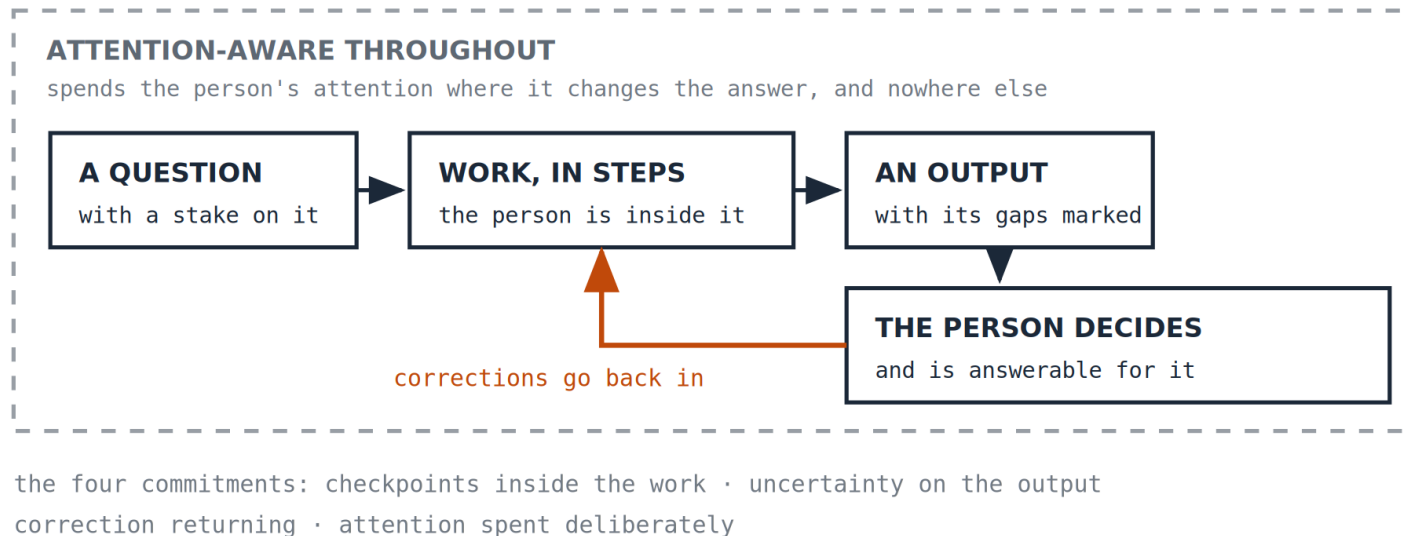
A tool that removes every occasion for the person to be wrong has removed the mechanism that made them worth consulting.

05 Both sides adapt.

The person learns the tool, which happens anyway. The tool also learns this person, this domain, this set of standing corrections.

Four commitments, drawn as one thing

Those five are a stance. These four are what you would have to put in the architecture, and they are one loop.



What honest uncertainty looks like

Of the four, this is the one you could build on Monday and the one almost nothing currently does.

ONE STATE OF THE EVIDENCE, TWO PRESENTATIONS

SEAMLESS



nothing here tells you where to push

HONEST



THE SOURCES DISAGREE HERE

NOTHING FOUND ON THIS AT ALL

now you know which two to check

same five lines, same model, different contract with the reader

Three tools that would follow

A RESEARCH TOOL

That leads with where the sources disagree rather than resolving the disagreement quietly on your behalf.

A WRITING TOOL

That asks a clarifying question before producing eight hundred words built on a guess about what you meant.

A DECISION TOOL

That states the assumptions its recommendation rests on, so you can disagree with an assumption rather than with a conclusion.

None of these is technically hard. All of them are commercially inconvenient, because each one makes the product look less capable than its competitor in a demonstration.

One of them exists

These commitments are not a wish list written by somebody who has never built anything. The author has a working prototype, a multi-agent research system, built to the commitments above.

SEVERAL AGENTS

With different jobs, rather than one model asked to do everything at once.

THE PERSON INSIDE

At the checkpoints, rather than reviewing at the end when the work is already shaped.

DISAGREEMENT KEPT

Where the agents conflict, the conflict is surfaced rather than averaged away.

It is a prototype and it is offered as one. What it demonstrates is not that the problem is solved, but that these commitments are things you can build with, rather than principles that evaporate on contact with an implementation.

What this does not claim

Short versions tend to overstate. Here is what this one is not saying.

✗ ~~AGI is impossible.~~

Not the argument, and not knowable. The argument is that it is the wrong thing to be aiming at, which is a question about goals rather than about feasibility.

✗ ~~This is a safer, more modest compromise.~~

It is not offered as a compromise. On this account it is the better goal on the technical merits, and the modest-sounding version is the one that actually works.

✗ ~~We already have this, it is called tools.~~

Today's tools optimise for engagement and task completion. One built to these commitments would optimise for the quality of the person's decision: a different target, and a different design. The test is what a system is built to maximise, not whether a human is present.

✗ ~~Here is the specification.~~

There isn't one. Five principles, four commitments and a prototype are a direction, not an engineering document, and the paper says so.

Who decides which is which

The honest hole in all of this. If the distinction is between augmenting a person and replacing them, somebody has to decide which is which, and that is not a technical judgement.

The same system is augmentation to the person using it and replacement to the person it was bought instead of.

There is no view from nowhere that settles it, and the people best placed to decide are rarely the people who carry the consequence. This paper does not solve that. It says the decision is a public one and should be made in public, which is a weaker answer than anybody wants and a more honest one than pretending the technology decides.

Three questions for any tool you are offered

1. WHAT DOES IT ASSUME ABOUT YOU?

That you have judgement worth preserving, or that you are an obstacle between the output and the outcome?

2. DOES IT EVER SAY IT DOES NOT KNOW?

If it has never once told you that the sources conflict or that it found nothing, it is not that it has always known.

3. WHAT CAN YOU DO NOW THAT YOU COULD NOT?

After a year of using it. If the honest answer is “produce more of what I used to produce”, the tool widened throughput. If it is “notice things I used to miss”, it widened you.

These work on any tool, including the ones the author builds.

Where this comes from

This is a short version of the constructive half of a working paper called AHI: The Case for Augmented Human Intelligence, which carries the full argument, the sources and the objections, including the ones that have not been answered.

The diagnosis it rests on is in the companion piece, Why Can't You Just Be Told?, and in the paper behind it, The Wisdom Gap.

The system amplifies. The human decides.

Working papers are revised when they are shown to be wrong, and what changed is kept on the record.

James Maconochie | Architecture & Attention

HUMILITAS, HUMANITAS, ET CURIOSITAS · SEMPER

Humility, humanity, and curiosity. Always.

