Agency: What it is and why it matters

Agency: What it is and why it matters

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

[ETA: I'm deprioritizing completing this sequence because it seems that other people are writing good similar stuff. In particular, see e.g. https://www.lesswrong.com/posts/kpPnReyBC54KESiSn/optimality-is-the-tiger-and-agents-are-its-teeth and https://www.lesswrong.com/posts/pdJQYxCy29d7qYZxG/agency-and-coherence]

This sequence explains my take on agency. I'm responding to claims that the standard arguments for AI risk have a gap, a missing answer to the question "why should we expect there to be agenty AIs optimizing for stuff? Especially the sort of unbounded optimization that instrumentally converges to pursuit of money and power."

This sequence is a pontoon bridge thrown across that gap.

I'm also responding to claims that there are coherent, plausible possible futures in which agent AGI (perhaps better described as APS-AI) isn't useful/powerful/incentivized, thanks to various tools that can do the various tasks better and cheaper. I think those futures are incoherent, or at least very implausible. *Agency is powerful.* For example, one conclusion I am arguing for is:

When it becomes possible to make human-level AI agents, said agents will be able to outcompete various human-tool hybrids prevalent at the time in every important competition (e.g. for money, power, knowledge, SOTA performance, control of the future lightcone...)

Another is:

We should expect Agency as Byproduct, i.e. expect some plausible training processes to produce agenty AIs even when their designers weren't explicitly aiming for that outcome.

I've had these ideas for about a year but never got around to turning them into rigorous research. Given my current priorities it looks like I might never do that, so instead I'm going to bang it out over a couple of weekends so it doesn't distract from my main work. :/ I won't be offended if you don't bother to read it.

Outline of this sequence:

1. P$_2$B: Plan to P$_2$B Better - LessWrong

2. Agents as P$_2$B chain reactions

3. Interlude: Agents as automobiles

4. Gradations of agency

5. Why agents are powerful, & other conclusions

Incomplete list of related literature and comments:

Frequent arguments about alignment - LessWrong (A comment in which Richard Ngo summarizes a common pattern of conversation about the risk from agenty AI vs. other sorts of AI risk)

Joe Carlsmith, drawing on writings from others, had 20% credence that AI agents won't be powerful enough relative to non-agents to be incentivised. I recommend reading the whole report, or at least the relevant sections on APS-AI and incentives to build it.

Eric Drexler's CAIS report (as summarized by Rohin Shah) argues basically that it should be much more than 20%. Richard Ngo's thoughts here.

Why You Shouldn't Be a Tool: The Power of Agency by Gwern. (OK, it seems to have a different title now, maybe it always did and I hallucinated this memory...) This essay, more than anything else, inspired my current views.

The Ground of Optimization by Alex Flint argues: "there is a specific class of intelligent systems — which we call optimizing systems — that are worthy of special attention and study due to their potential to reshape the world. The set of optimizing systems is smaller than the set of all AI services, but larger than the set of goal-directed agentic systems."

Yudkowsky and Ngo conversation (especially as summarized by Nate Soares) seems to be arguing for something similar to Alex -- I imagine Yudkowsky would say that by focusing on agency I'm missing the forest for the trees: there is a broader class of systems (optimizers? consequentialists? makers-of-plans-that-lase?) of which agents are a special case, and it's this broader class that has the interesting and powerful and scary properties. I think this is probably right but my brain is not yet galaxy enough to grok it; I'm going to defy EY's advice and keep thinking about the trees for now. I look forward to eventually stepping back and trying to see the forest.

Thanks to various people, mostly at and around CLR, for conversations that shaped my views on this subject. Thanks especially to Ramana Kumar whose contributions were the greatest.

P$_2$B: Plan to P$_2$B Better

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

tl;dr: *Most good plans involve taking steps to make better plans. Making better plans is* ***the*** *convergent instrumental goal, of which all familiar convergent instrumental goals are an instance. This is key to understanding what agency is and why it is powerful.*

Planning means using a world model to predict the consequences of various courses of actions one could take, and taking actions that have good predicted consequences. (We think of this with the handle "doing things for reasons," though we acknowledge this may be an idiosyncratic use of "reasons.")

We take "planning" to include things that are relevantly similar to this procedure, such as following a bag of heuristics that approximates it. We're also including actually following the plans, in what might more clunkily be called "planning-acting."

Planning, in this broad sense, seems essential to the kind of goal-directed, consequential, agent-like intelligence that we expect to be highly impactful. This sequence explains why.

One Convergent Instrumental Goal to Rule them All

Consider the maxim

> *"make there be more and/or better planning towards your goal."*

This section argues that all the classic convergent instrumental goals are special cases of this maxim.

To flesh this out a little, here are some categories of ways to follow the maxim. Remember that a planner is typically close (in terms of what it might affect via action) to at least one planner – itself – so these directions can typically be applied in the first case to the planner itself.

- Make the planners with your goal **better at planning**. For example, get them new relevant data to work with[1], get them to run faster or more effective algorithms, build protections against value drift, etc.
- Make the planners with your goal have **better options**. For example, move them to better locations, get them more resources, get them more power or a greater number of options to select from, have them take steps in an object-level plan towards the goal.
- Make there be **more planners** with your goal. For example, keep yourself running and aligned with your goal, acquire delegates and subordinates, convince followers and converts, build successors.

Reviewing Omohundro's “The Basic AI Drives” and Bostrom’s “The Superintelligent Will,” we extract a list of convergent instrumental goals, and find that they are all instances of the maxim “make there be more/better planners for the current goal:”

- **Self-preservation / self-protection:**
    - Make there be more planners that have your goals, focusing on reusing the existing planner, that is, preventing its destruction.
- **Self-improvement:**
    - Make there be better planners with your goals, focusing on making the existing one better.
- **Resource Acquisition:**
    - Make there be better planners with your goals, focusing on making the existing one able to take more effective actions.
- **Goal-content Integrity:**
    - Make there be more planners that have your goals, focusing on ensuring the existing planners that have your goals keep those goals and avoid them being changed.
- **Resource-use Efficiency:**
    - Same as self-improvement
- **Cognitive Enhancement:**
    - Same as self-improvement
- **Creativity:**
    - Same as self-improvement
- **Technological Perfection:**
    - Same as resource acquisition and/or self-improvement
- **Rationality:**
    - Omohundro takes this to be something like “make the utility function explicit” along with “maximize expected utility.”
    - Thus it is similar to goal-content integrity and self-improvement.
- **Utility-function preservation:**
    - Similar to goal-content integrity.
- **Prevent counterfeit utility:**
    - Essentially this is avoiding wireheading. Omohundro: “An important class of vulnerabilities arises when the subsystems for measuring utility become corrupted.”
    - Thus it is similar to goal-content integrity.

Seeking a concise, memorable-yet-accurate name for this maxim, this convergent-instrumental-goal-to-rule-them-all, we settled on P_2B:

P_2B ≔ Plan to P_2B Better

This name emphasizes the recursive, feedback-loopy aspect of the phenomenon, which is only implicit in the idea of “better plans.”

Why P_2B Works

There are several ways for planning to be ineffective, such as an inaccurate or unwieldy world model, a limited selection of actions to choose from, or inefficient use of time or other resources in predicting or assessing consequences. But often planners can and will address these issues: **planning is self-correcting**, **thanks to $\mathbf{P_2B}$**. A planner that didn’t recognize the importance of P_2B, or was unable to do it for some reason, would not be self-correcting.

Instrumental goals are about passing the buck: if you are a planner, and you can’t achieve your final goal with a single obvious action (or sequence of actions), you can instead pass the buck to something else, typically your future self.[2] There will often be obvious available actions that put the receiver of the buck “closer” to achieving the final goal than you.[3]

P_2B is what it means to generically pass the buck, closing some distance along the way. The “better” in P_2B means being closer to the goal and/or generally able to close distance faster. Convergent instrumental goals are ways to close distance that work for almost any final goal, hence they are instances of P_2B.

Objections & Nuances

Isn’t this trivial?

One might complain that we’ve defined P_2B broadly enough that our claim about it being **the** convergent instrumental goal is trivial — true by definition since we defined “planner” and “better” so broadly. **Fair enough**; the reason we are doing this is because we think it's a useful framing/foundation for answering questions about agency, not because we think it is important or interesting on its own. We agree that the more detailed taxonomies of instrumentally convergent goals are useful. We just think it is also useful to have this unified frame. We intend to write subsequent posts making use of this frame.

Most agents aren’t planners though?

Yes they are — remember, we said above that we are defining “planner” broadly to include relevantly similar algorithms/procedures. Let’s flesh that idea out a bit more…

We think it’s OK to talk loosely about families of algorithms. When we say “planning,” for example, we are gesturing at a vague cluster of algorithms that has “for each action, imagine the expected consequences of that action, then evaluate how good those consequences are, then pick the action that had the best expected consequences” as a central example.

We are *not* saying that every planning algorithm must be exactly of that form. Examining exactly where the boundaries of these concepts lie is an interesting and potentially valuable rabbit hole that we don’t feel the need to go down yet.

One thing we do wish to say is that we intend to include algorithms which *behave similarly* to the paradigmatic planning algorithm mentioned above. One easy way to generate algorithms like this is by automating bits of the process with heuristics. For example, maybe instead of calculating the expected consequences of every action all the time, the algorithm has a bag of heuristics that tell it when to calculate and when to not bother (and what to do instead) and the bag of heuristics tends to yield similar results for less computational expense, at least in some relevant class of environments.[4]

(We haven't defined "agents" yet, but you can probably guess from what we've said that our definition is going to resemble Dennett's Intentional Stance.)

What about the procrastination paradox?

A planner that "P_2Bs forever", without ever taking "object-level" actions in plans that aren't about making better future plans/planners, won't be very effective at achieving its goal. But P_2B is not the only strategy a planner should pursue — we have only said that P_2B is the convergent **instrumental** goal. Whenever there are obvious actions that directly lead towards the goal, a planner should take them instead.

The danger of taking instrumental actions forever can show up in some toy decision problems. However, in realistic cases and for realistic planners, this is not so much of an issue — one can pursue convergent instrumental goals without ceasing to keep an eye out for opportunities to achieve terminal goals. Nevertheless, due to the automation-of-bits-of-the-core-algorithm phenomenon described above, it's not uncommon for agents to end up pursuing P_2B as a terminal goal, or even pursuing sub-sub-subgoals of P_2B such as "acquire money" as final goals. As Richard Ngo pointed out, we should expect mesa-optimizers to develop *terminal* goals for power, survival, learning, etc. because such things are useful in a wide range of environments and therefore probably useful in the particular environment they are being trained in.

Footnotes

[1] This was an "aha" moment for me: *Even such everyday actions as "briefly glance up from your phone so you can see where you are going when walking through a building" are instances of following this maxim!* You are looking up from your phone so that you can acquire more relevant data (the location of the door, the location of the door handle, etc.) for your immediate-future-self to make use of. Your immediate-future-self will have a slightly better world-model as a result, and thus be better than you at making plans. In particular, your immediate future self will be able, e.g., to choose the correct moment & location to grab the door handle, by contrast with your present self who is looking at Twitter and does not know where to grab. ↩

[2] This phrasing makes it sound like your goal is binary and permanent, either achieved or not. For more typical goals, which look more like utility functions, we think the same point would apply but would be more unwieldy to state. ↩

[3] To make this analogy to covering distance from a target more precise, consider something like the edit distance between the world as it is and any variation satisfying one's final goal. The world is always changing, but the fraction of changes that are reducing this edit distance increases when there are more capable and effective planners with that goal. ↩

[4] Or perhaps even it's heuristics all the way down, but it's a sophisticated bag of heuristics that behaves *as if* it were following the calculate-expected-consequences-then-pick-the-best procedure, at least in some relevant class of environments. Note that we have the intuition that, generally speaking, substituting heuristics for bits of

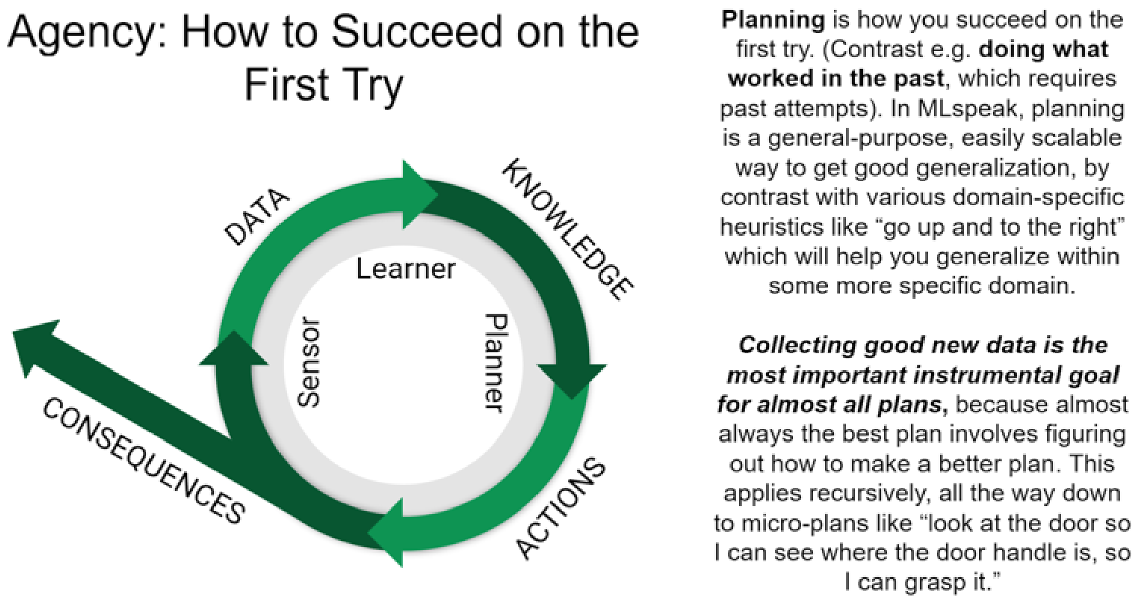
the core algorithm risks increasing “brittleness”/”narrowness,” i.e., problematically reducing the range of environments in which the system behaves like a planner. ↩

Agents as P_2B Chain Reactions

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

tl;dr: *Sometimes planners successfully P_2B, kicking off a self-sustaining chain reaction / feedback loop of better and better plans (made possible by better and better world-models, more and more resources, etc.) Whereas fire takes a concentration of heat as input and produces a greater concentration of heat as output, agents take a concentration of convergent instrumental resources (e.g. data, money, power) as input and produce a greater concentration as output.*

Previously we described the Convergent-Instrumental-Goal-to-rule-them-all, P2B: Plan to P2B Better. The most common way to P_2B is to plan to plan again, in the immediate future, but with some new relevant piece of data. For example, suppose I am walking somewhere. I look up from my phone to glance at the door. Why? Because without data about the location of the doorknob, I can’t guide my hands to open the door. This sort of thing doesn’t just happen on short timescales, though; humans (and all realistic agents, I’d wager) are hierarchical planners and medium and long-term plans usually involve acquiring new relevant data as well.


Agency: How to Succeed on the First Try

Planning is how you succeed on the first try. (Contrast e.g. **doing what worked in the past**, which requires past attempts). In MLspeak, planning is a general-purpose, easily scalable way to get good generalization, by contrast with various domain-specific heuristics like “go up and to the right” which will help you generalize within some more specific domain.

Collecting good new data is the most important instrumental goal for almost all plans, because almost always the best plan involves figuring out how to make a better plan. This applies recursively, all the way down to micro-plans like “look at the door so I can see where the door handle is, so I can grasp it.”

This image just talks about collecting good new data, but there are other important convergent instrumental goals too. Such as staying alive, propagating your goals to other agents, and acquiring money. One could modify the diagram to include these other feedback loops as well — more data *and money and power* being recruited by sensor-learner-planners to acquire more data *and money and power*. It’s not that all of these metrics will go up with every cycle around the loop; it’s that some goals/resources are “instrumentally convergent” in that they tend to show up frequently in most real-world P_2B loops.

Agents, I say, are *P_2B chain reactions / P_2B feedback loops.* They are what happens when planners successfully P_2B. Whereas fire is a chain reaction that inputs heat + fuel + oxygen and outputs more heat (and often more oxygen and fuel too, as it expands to envelop more such) agents are chain reactions that input sensor-learner-planners with some amount of data, knowledge, power, money, etc. and output more, better sensor-learner-planners with greater amounts of data, knowledge, power, money, etc.

(I don’t I think this definition fully captures our intuitive concept of agency. Rather, P_2B chain reactions seem like a big deal, an important concept worth talking about, and close enough to our intuitive concept of agency that I’m going to appropriate the term until someone pushes back.)

Already you might be able to guess why I think agents are powerful:

- Consider how fire is a much bigger deal than baking-soda-vinegar. General/robust feedback loops are *way way* better/more important than narrow/brittle ones, and *sensor-learner-planner* makes for a P_2B loop that is pretty damn general.
- Ok, so we are talking about a kind of chain reaction that takes concentrations of knowledge, power, money, etc. and makes them bigger? That sure sounds like it’ll be relevant to discussions of who’s going to win important competitions over market share, political power, and control of the future!

The next post in this sequence will address the following questions, and more:

> OK, so does any feedback loop of accumulating convergent instrumental resources count as an agent then? Presumably not, presumably it has to be resulting from some sort of planning to count as a P2B loop. Earlier you said planning is a family of algorithms… Say more about where the borders of this concept are please!

The post after that will answer the Big Questions:

> Why is agency powerful? Why should we expect agent AGIs to outcompete human+tool combos for control of the future? Etc.

Interlude: Agents as Automobiles

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

I ended up writing a long rant about agency in my review of Joe Carlsmith's report on x-risk from power-seeking AI. I've polished the rant a bit and posted it into this sequence. The central analogy between APS-AI and self-propelled machines ("Auto-mobiles") is a fun one, and I suspect the analogy runs much deeper than I've explored so far.

For context, the question being discussed is whether we should "expect incentives to push relevant actors to build agentic planning and strategically aware systems [APS systems] in particular, once doing is possible and financially feasible."

Joe says 80% yes, 20% no:

"The 20% on false, here, comes centrally from the possibility that the combination of agentic planning and strategic awareness isn't actually that useful or necessary for many tasks -- including tasks that intuitively seem like they would require it (I'm wary, here, of relying too heavily on my "of course task X requires Y" intuitions). For example, perhaps such tasks will mostly be performed using collections of modular/highly specialized systems that don't together constitute an APS system; and/or using neural networks that aren't, in the predictively relevant sense sketched in 2.1.2-3, agentic planning and strategically aware. (To be clear: I expect non-APS systems to play a key role in the economy regardless; in the scenarios where (2) is false, though, they're basically the only game in town.)"

I agree that "of course X requires Y" intuitions have been wrong in the past and also that evidence from how nature solved the problem in humans and nonhuman animals will not necessarily generalize to artificial intelligence. However:

1. Beware isolated demands for rigor. Imagine someone in 1950 saying "Some people thought battleships would beat carriers. Others thought that the entire war would be won from the air. Predicting the future is hard; we shouldn't be confident. Therefore, we shouldn't assign more than 90% credence to the claim that powerful, portable computers (assuming we figure out how to build them) will be militarily useful, e.g. in weapon guidance systems or submarine sensor suites. Maybe it'll turn out that it's cheaper and more effective to just use humans, or bio-engineered dogs, or whatever. Or maybe there'll be anti-computer weapons that render them useless. Who knows. The future is hard to predict." This is what Joe-with-skeptic-hat-on sounds like to me. Battleships vs. carriers was a relatively hard prediction problem; whether computers would be militarily useful was an easy one. I claim it is obvious that APS systems will be powerful and useful for some important niches, just like how it was obvious in 1950 that computers would have at least a few important military applications.
2. To drive this point home, let me take the reasons Joe gave for skepticism and line-by-line mimic them with a historical analogy to self-propelled machines, i.e. "automotives." Left-hand column is entirely quotes from the report.

Skeptic about APS systems	**Skeptic about self-propelled machines**
Many tasks -- for example, translating languages, classifying proteins, predicting human responses, and so	Many tasks — for example, raising and lowering people within a building, or transporting slag from the mine to the deposit, or transporting water from the source to the home — don't seem to require automotives. Instead, an engine can be fixed in one location to power a system of pulleys or conveyor belts, or pump liquid through pipes. Perhaps all or most of the tasks involved

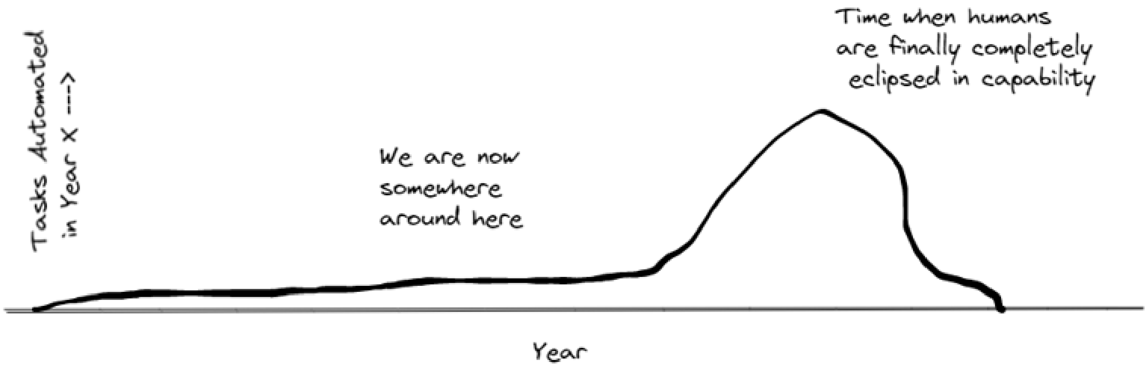
forth -- don’t seem to require agentic planning and strategic awareness, at least at current levels of performance. Perhaps all or most of the tasks involved in automating advanced capabilities will be this way.	in automating our transportation system will be this way.
In many contexts (for example, factory workers), there are benefits to specialization; and highly specialized systems may have less need for agentic planning and strategic awareness (though there’s still a question of the planning and strategic awareness that specialized systems in combination might exhibit).	In many contexts, there are benefits to specialization. An engine which is fixed to one place (a) does not waste energy moving its own bulk around, (b) can be specialized in power output, duration, etc. to the task at hand, (c) need not be designed with weight as a constraint, and thus can have more reliability and power at less expense.
Current AI systems are, I think, some combination of non-agentic-planning and strategically unaware. Some of this is clearly a function of what we are currently able to build, but it may also be a clue as to what type of systems will be most economically important in future.	Current engines are not automotives. Some of this is clearly a function of what we are currently able to build (our steam engines are too heavy and weak to move themselves) but it may also be a clue as to what type of systems will be most economically important in the future.
To the extent that agentic planning and strategic awareness create risks of the type I discuss below, this	To the extent that self-propelled machines may create risks of “crashes,” this might incentivize focus on other types of systems (and I would add that a fixed-in-place engine seems inherently safer than a careening monstrosity of iron and coal!) To the extent that self-propelled machines may enable some countries to

might incentivize focus on other types of systems.	invade other countries more easily, e.g. by letting them mobilize their armies and deploy to the border within days by riding “trains,” and perhaps even to cross trench lines with bulletproof “tanks,” this threat to world peace and the delicate balance of power that maintains it might incentivise focus on other types of transportation systems. *[Historical note: The existence of trains was one of the contributing causes of World War One. See e.g. Railways and the mobilisation for war in 1914 | The National Archives.]*
Plan-based agency and strategic awareness may constitute or correlate with properties that ground moral concern for the AI system itself (though not all actors will treat concerns about the moral status of AI systems with equal weight; and considerations of this type could be ignored on a widespread scale).	OK, I admit that it is much more plausible that people will care for the welfare of APS-AI than for the welfare of cars/trains. However I don’t think this matters very much so I won’t linger on this point.

3. There are plenty of cases where human “of course task X requires Y” intuitions turned out to be basically correct. (e.g. self-driving cars need to be able to pathfind and recognize images, image-recognizers have circuits that seem to be doing line detection, tree search works great for board game AIs, automating warehouses turned out to involve robots that move around rather than a system of conveyor belts, automating cruise missiles turned out to *not* involve having humans in the loop steering them… I could go on like this forever. I’m deliberately picking “hard cases” where a smart skeptic could plausibly have persuaded the author to doubt their intuitions that X requires Y, as opposed to cases where such a skeptic would have been laughed out of the room.)

4. There’s a selection effect that biases us towards thinking our intuition about these things is worse than it is:

- Cases where our intuition about is incorrect are cases where it turns out there is an easier way, a shortcut. For example, chess AI just doing loads of really fast tree search instead of the more flexible, open-ended strategic reasoning some people maybe thought chess would require.
- If the history of AIs-surpassing-humans-at-tasks looks like this:



- Then we should expect the left tail to contain a disproportionate percentage of the cases where there is a shortcut. Cases where there is no shortcut will be clumped over on the right.

4. More important than all of the above: As Gwern pointed out, *it sure does seem like some of the tasks some of us will want to automate are agency tasks*, tasks such that anything which performs them is by definition an agent. Tasks like “gather data, use it to learn a general-purpose model of the world, use that to make a plan for how to achieve X, carry out the plan.”

5. Finally, and perhaps most importantly: **We don’t have to go just on intuition and historical analogy. We have models of agency, planning, strategic awareness, etc. that tell us how it works and why it is so useful for so many things.** [This sequence is my attempt to articulate my model.]

Many thanks to Joe Carlsmith for his excellent report and for conversing with me at length about it.

Gradations of Agency

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

[Epistemic status: This post written hastily, for Blog Post Day.]

I've de-prioritised this sequence due to other recent posts which cover a lot of the same ground. Moreover I begin to suspect that the theory of agency I'm building is merely the glowing brain and that Yudkowsky-Veedrac's is the cosmic galaxy brain. Perhaps mine can be a useful stepping-stone.

Consider the messy and large real world, with creatures reproducing, competing, and evolving. The behavior of each entity is controlled by some sort of cognition/computation—some sort of algorithm.

Inspired by something Dan Dennett wrote (h/t Ramana Kumar), I propose the following loose hierarchy of algorithm families. As we move up the hierarchy, things get more complicated and computationally expensive, but also more powerful and general.

Level 1: *Do what worked in the past for your ancestors.* You have some "input channel" or "senses" and then you respond to it somehow, doing different things in response to different inputs.

Example: You swim towards warmth and away from cold and hot. You run from red things and towards green things.

This is more complex and computationally expensive, but (if done right) more powerful and general than algorithms which aren't environment-responsive: There are many situations, many niches in the environment, where being able to sense and adapt quickly helps you survive and thrive.

Level 2: *Do what worked in the past for you.* How you respond to inputs is itself responsive to inputs; you learn over the course of your life. This typically involves some sort of memory and a special input channel for positive and negative reward.

Example: You start off reacting randomly to inputs, but you learn to run from red things and towards green things because when you ran towards red things you got negative reward and when you ran towards green things you got positive reward.

This is more complex and computationally expensive, but (if done right) more powerful and general:

- For a particular ecological niche, you & your relatives will reproduce and fill the niche more quickly than rivals from Level 1. This is because your rivals in some sense keep doing the same thing until they die, and will thus only "learn" to exploit the niche by gradual natural selection.
- Also, there are some niches that Level 1 creatures simply cannot survive in, that you can. Example: In Pond X, the color of food and the color of predators constantly changes on timescales of about a lifetime, too fast for a population of Level 1 creatures to adapt.

Level 3: *Do what worked in the past for things similar to you in similar situations.* You have some ability to judge similarity, and in particular to notice and remember when your current situation is similar to the situation of something else you saw, something you classify as similar to you. Armed with this ability you can learn not just from your own experience, but from the experience of others—you can identify successful others and imitate them.

This is more complex and computationally expensive, but (if done right) more powerful and general:

- For a particular ecological niche, you & your relatives will reproduce and fill the niche more quickly than rivals from level 2. This is because you can learn from the experiences of others while level 2 creatures have to experience things themselves.
- Also, there are some niches that level 2 creatures simply cannot survive in, that you can. For example, niches in which certain behaviors in certain situations are deadly.

Level 4: *Do what worked in your head when you imagined various plans*. Whereas with Level 3 you processed your memories into the sort of world-model that allowed you to ask "Have I seen anyone similar to me in a similar situation before? What did they do—did it work for them?" now you are processing your memories into the sort of gearsy world-model that allows you to imagine the consequences of different hypothetical actions and pick the action has the best imagined consequences. This includes level 3 thinking as a special case.

This is more complex and computationally expensive than level 3, but (if done right) more powerful and general. (To save space, spelling out why is left as an exercise for the reader.)

Level 5: *Cognitive algorithms learned from experience*. Remember, the levels we've been describing so far are algorithm families. There are lots of choices to be made within each family about exactly how to do it—e.g. How many possible actions do you consider, and how do you choose which ones? How long do you spend imagining the outcomes of each possible action? How should the answers to the previous questions depend on your situation? Since I didn't say otherwise, assume that in previous levels the answers to these questions are fixed and hard-coded; now, at Level 5, we imagine that they are variable parameters you can adjust depending on the situation and which you can learn over the course of your life.

Remember that realistically you won't be just doing one level of cognition all the time; that's massively computationally wasteful. Instead you'll probably be doing some sort of meta-algorithm that switches between levels as needed.

(For example, there's no need to imagine the consequences of all my available actions and pick the best plan if I'm doing something extremely similar to what I did in the past, like brush my teeth; I can handle those situations "on autopilot." So maybe the first few times I brush my teeth I do it the computationally expensive way, but I quickly learn simple algorithms to "automate" the process and then thenceforth when I notice I'm in the teeth-brushing situation I switch on those algorithms and let the imitation and planning parts of my brain relax or think about other things.)

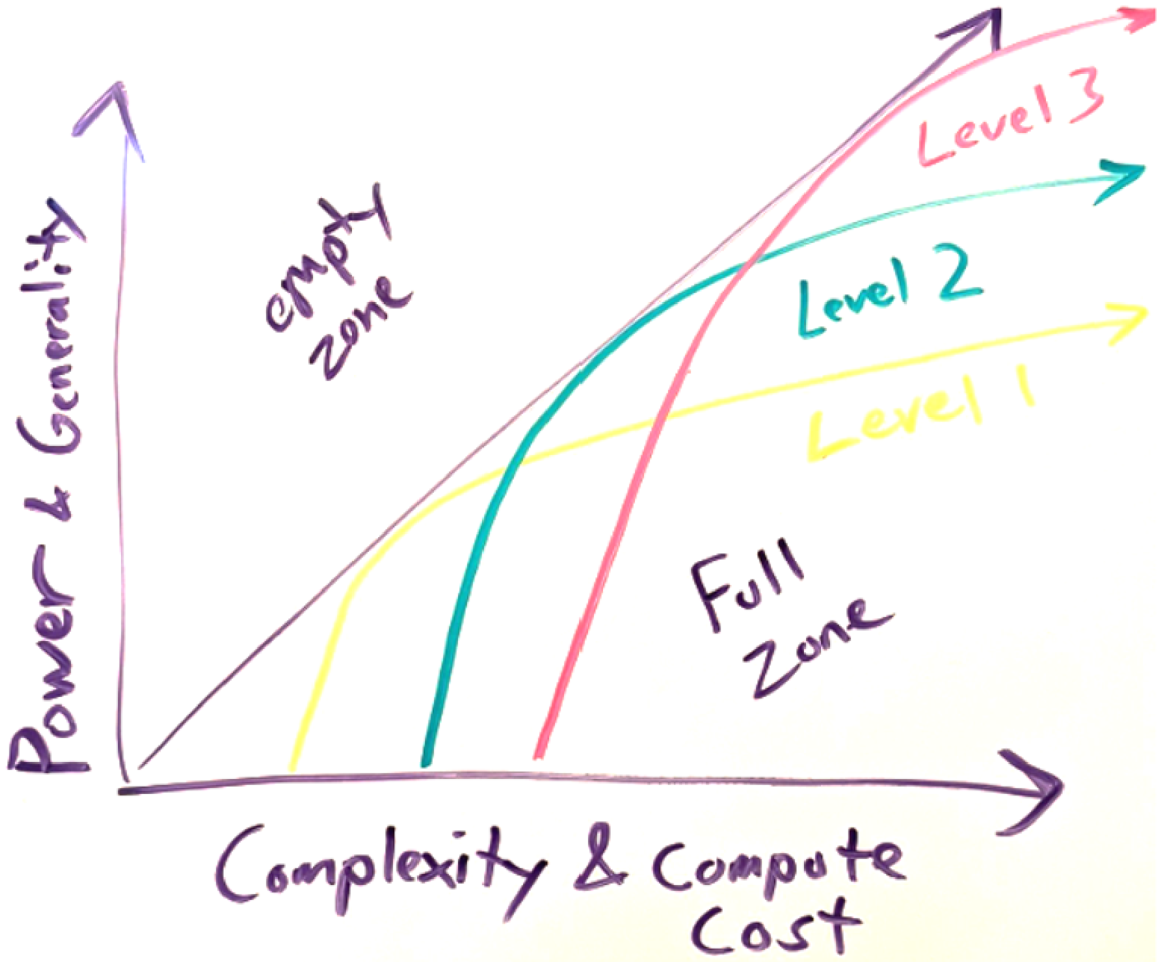
Given that this is what you'll be doing, *whether to delegate to some cheaper/faster process, and which one to delegate to,* is itself a difficult question that is best answered with a learned, rather than hard-coded, algorithm. Hence the importance of Level 5.

Level 6: *Cognitive algorithms learned from imitation and imagined experience.* Whereas in Level 5 the meta-cognitive process that chose details of your algorithm was simple: "think in ways that seem to have worked in the past;" now you are able to apply your imitation and planning algorithms on this meta level: "think in ways that worked for others" and "think in ways that you predict will work better than the alternatives."

…I'll stop there. I'm not claiming that there are no levels above 6; nor am I claiming that this way of carving up the space of algorithms is the best way. Here are my conjectures:

1. We can imagine a graph with algorithmic complexity + computational cost on the x-axis, and power level (I haven't defined this, but note that each of the levels both *learns more quickly* and *can succeed in a wider variety of situations* compared to the previous levels) on the y-axis. The different algorithm families can be depicted on the graph like so. (Each instance of a family lies somewhere below that family's curve.) A process that optimizes for power+generality while searching deeper and deeper into

the y-axis (such as a series of neural net training runs with larger and larger neural nets, or a single training run perhaps) will tend to progress up some sort of hierarchy similar to the six-level hierarchy I've described.



2. Intuitively, each level seems "more agenty" than the previous levels. Perhaps in the infinite limit, you get something like a Bayesian expected utility maximizer for which utility = power-and-generality-as-defined-by-the-graph. Could it instead be that if we keep exploring the diagram farther to the right, the trend towards agency will reverse? Maybe, but I would be quite surprised.
3. Each level seems to be capable of creating and sustaining more powerful, more robust P_2B loops than the previous levels. (Arguably levels 1, 2, 3 aren't really P_2B loops because they don't involve planning… but I don't want to say that exactly because e.g. level 3 involves imitating others and so that probably ends up accumulating resources and knowledge and other convergent instrumental resources, and generally behaving very similarly to a planner at least in some environments. And likewise level 2 can sorta crappily imitate level 3… perhaps I should say that the key thing is "convergent instrumental resource feedback loops" and that thinking about P_2B is a way to see why such things happen and why the sufficiently big ones tend to have planners sitting in the middle directing them, and the really really big ones tend to have level 6+ algorithms sitting in the middle directing them…

The next and probably last post in this sequence ties it all together & answers the big questions about agency.

Why agents are powerful

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

[Written for Blog Post Day. Not super happy with it, it’s too rambly and long, but I’m glad it exists.]

Here are some questions I think this theory of agency can answer:

- What are agents?
- Why should we expect AI agents to be useful for various important tasks?
- Why should we think agentic mesa-optimizers may arise for some data+reward-signals that weren’t designed explicitly to produce them?
- Why should we think humans+AI tools will eventually be outcompeted by AI agents?
- Why should we expect, on priors, mildly superhuman AI agents to be powerful enough to take over the world, if they wanted to?

What are agents?

Earlier in “Agents as P_2B chain reactions” I defined agents as successful P_2B feedback loops—a learner algorithm and a planner algorithm hooked up together, the planner plans to plan better and thus outputs actions that result in getting more useful data into the learner, more resources for the planner to plan with… then the process repeats… Like how “fire” is a successful feedback loop in which heat+plasma+fuel come together to produce more heat+plasma, and envelop more fuel, if any is nearby.

Then I asked: Is the planner algorithm strictly necessary here? Could you get something similar to a P_2B feedback loop, but built around something other than a planner? I explored this question indirectly in “Gradations of Agency.” The answer seems to be “sorta.” You can have e.g. a Level 3 system that doesn’t do any planning and yet behaves remarkably similar to a P_2B feedback loop in many contexts, due to imitating the success of others. Such a system would tend to underperform P_2B in contexts where there aren’t good examples to imitate.

We could stick to the original definition and say: Agents are P_2B feedback loops, so, level 4 and above.

But I think it’s probably better to be a bit more galaxy brain and say: the key thing is *convergent instrumental resource feedback loops* and that thinking about P_2B is a stepping stone which helps us see why such phenomena exist and why the really big ones tend to have planners sitting in the middle directing them, and the really *really* big ones tend to have level 6+ algorithms sitting in the middle directing them…

By analogy, heat+plasma+wood+oxygenl chain reactions aren’t the key thing. The key thing is *chain reactions that convert some sort of fuel into thermal and kinetic energy plus more of the chain reaction.* Heat+plasma+wood+oxygen is a particularly important one, due to the ease of creating it and the abundance of fuel for it on Earth, but there’s also e.g. baking soda + vinegar and neutrons+uranium. And also self-replicating nanobots + pretty much anything.

Similarly: The important thing for understanding agency is understanding that *the world contains various self-sustaining chain reactions that feed off instrumental resources like data, money, political power, etc. and spread to acquire more such resources.* Level 4 P_2B loops are a particularly important, powerful, and widespread instance of this type, analogous to fire, but there are less powerful/important things (such as Level 3, Level 2…) and also more powerful/important things (Level 6 and above?).

Why should we expect AI agents to be useful for various important tasks?

Why should we think agentic mesa-optimizers may arise for some data+reward-signals that weren't designed explicitly to produce them?

Both questions have the same answer. *Agency works*; it's powerful and general. Recalling this graph and the conjectures from last post, I claim that as you crank up the "how fast does this thing learn, how high a score does it get, in how large and diverse a range of environments?" dial, you ascend the hierarchy until you get agents. You can get powerful non-agents but only by doing something clever and costly, only by selecting against agency or something similar. If you just select/optimize for power/generality/etc. then you get agents. (In other words, for sufficiently large x, the most powerful/general/etc. algorithm with x complexity and compute cost is an agent.)

The reason agency works basically boils down to the claims made way back in P_2B: Plan to P_2B Better. Accumulating knowledge, power, resources, etc. and using them to accumulate even more such things, in an epic feedback loop, until you can convert that pile of knowledge+power+resources into X, is a great way to achieve X, for a very broad range of X. In fact it's the best/easiest way, for a broad range of definitions of best/easiest. These piles of resources don't accumulate by themselves; they accrue around algorithms, algorithms such as the learner (level 2), the imitator (level 3), the planner (level 4) … and they accrue faster around higher-level algorithms.

(Insofar as there is a better strategy for achieving X than the above, a level 4+ algorithm sitting on top of a sufficiently large pile of knowledge+power+etc. will realize this and pursue that better strategy instead, self-modifying if necessary.)

The analogy to automobiles is relevant here. Let X be some transportation job you want to automate—such-and-such needs to be moved from A to B, quickly cheaply and safely. For a very broad range of X, the best strategy is to use some sort of auto-mobile: a machine that moves itself. As opposed to e.g. a series of conveyor belts, or vacuum pipes, or a bucket brigade of cranes. While there are gains from specialization, it's much better to build "general automobiles" that are modular — they have a compartment for holding the passengers/cargo, and then the rest of the machine is dedicated to quickly, cheaply, and safely moving the entire assemblage. The strategy is to move the entire assemblage from pretty-much-A to pretty-much-B, with a short step at the beginning and end for loading and unloading. Some kinds of automobiles rely on the cargo for motive power, e.g. a dirtbike relies on the physical strength of the rider and their shifting center of mass to steer and pop wheelies over curbs etc. But most, and especially the more powerful ones (the ones required for the more difficult X), treat the cargo/passengers as useless dead weight, because they simply aren't able to contribute much except at the beginning and end (loading and unloading).

Similarly, for a wide range of goals (metrics X that you want to go up), if X is sufficiently challenging, the optimal strategy for achieving X is to build a modular, agentic system that has a compartment for holding X, and that mostly ignores X "during the journey." X is stored away safely somewhere while the agent accumulates more data, resources, power, etc. so that it can accumulate even more of the same… and then eventually it has so much of all these things that it can just "directly achieve the goal" slash "switch to exploitation of X." Then X is "unloaded" and starts actually influencing how the system behaves.

All of this is an empirical claim about the structure of our world. There are possible worlds where it isn't true, just as there are possible worlds where building automobiles was more expensive and less effective than putting conveyor belts everywhere. But it seems clear that we aren't in one of those worlds.

Why should we think humans+AI tools will eventually be outcompeted by AI agents?

So we've got these AI agents, algorithms that repeatedly P_2B, accumulating piles of knowledge+power+money around them. Great. Humans and human corporations also do that. And as AI capabilities advance, and AI agents get more powerful, AI tools will also get more powerful. Perhaps humans + AI tools will be able to always stay one step ahead of AI agents.

I don't think so. AGI agents can use the AI tools too.

Recall the automobiles analogy: For sufficiently powerful automobiles doing sufficiently difficult transportation jobs, it's counterproductive try to make use of the human passenger's muscles. The passenger is dead weight. Similarly, for agency: Agents are feedback loops; an agent without a human-in-the-loop is an AI agent, an agent with a human-in-the-loop (can be) a human agent. (Depends on which part of the loop the human does.) Whatever it is that the human is doing, can be done faster and better by AI. The only way for a human-in-the-loop system to be competitive, in the limit of increased powerfulness, is for the human to stay out of the way and just come along for the ride. Such a system is a human-aligned AGI agent, not a human+tool hybrid agent.

(Continuing the analogy, we could argue that human-aligned AGI agents will eventually be outcompeted by unaligned AGI agents in the same way that aircraft steered by computers but still with a human pilot present in the cockpit will eventually be outcompeted by drones that don't need a cockpit. However, hopefully we won't ever let competition get that fierce; hopefully once we have human-aligned AGI agents they will be able to prevent the creation of unaligned AGI thenceforth.)

Why should we expect, *on priors,* mildly superhuman AI agents to be powerful enough

to take over the world, if they wanted to?

I have a lot to say on the subject of how easy it would be for mildly superhuman AI agents to take over the world. I've already said some of it already. In this section I'll give a brief, abstract argument from analogy:

Suppose there was a different and exotic chemical reaction called "Magefyre." Magefyre is purple instead of orange. Whereas regular fire takes wood, oil, etc. plus oxygen and produces more of itself, sending out sparks in all directions, etc., Magefyre does the same thing but is slightly "better" … more heat, more sparks, the sparks fly farther, more diverse kinds of fuel can catch Magefyre than catch fire, the fuel burns more quickly…

If a city is in the process of burning down due to regular fire, and a wizard starts a small Magefyre in some corner of the city, we should expect the Magefyre to eventually "take over" and end up burning down more of the city than the regular fire. The exceptions would be if the whole city burns down quickly before Magefyre can grow big, or if the Magefyre gets an unlucky start and finds itself surrounded by large amounts of regular fire that use up all the fuel in the vicinity. (The analogy to COVID strains is also helpful here; consider how "better" strains came along and quickly became dominant, even though they started with orders of magnitude fewer hosts.)

Thus, my answer: There is some core metric of "how good are you at quickly and robustly accumulating convergent instrumental resources you can use to repeat this process…" In other words, how powerful of an agent are you? There is an analogous metric for different kinds of fire, and for different strains of a virus.

Exceptions excluded, in the long run, small differences in agents in this metric will translate into large differences in outcome; the most powerful agent will win. And the long run here isn't actually that long; it will feel surprisingly short, just as it intuitively feels like Omicron Strain took over quickly from Delta.

Presumably this metric is an aggregate of different sub-metrics. For example if Magefyre is sufficiently superior to fire in some ways, it can still be overall better than fire even if it is worse in other ways. With agency, we should note that we could have mildly superhuman AI agents (in the relevant metric) that are still clearly subhuman in some important skills.

Example: Maybe the first "mildly superhuman" AGI agent uses 100x more compute than the human brain per second, and also requires 100x more experience/data to learn anything. Also, it can't do image processing or motor control. So it's significantly subhuman in these ways. But the AI lab that built it has enough compute to run 100,000 copies in parallel, and they have the key advantage that they can share their experience/learnings with each other, and so they end up quickly accumulating more know-how in relevant areas like reasoning, strategy, persuasion, etc. than any human ever, and in particular more than the groups of humans who currently oppose them; they then leverage this know-how to neutralize their immediate opponents, consolidate their power, and accumulate more compute + human patsies…

Wrapping up

What do I think of this theory, overall? Ehh, it’s definitely a beta version at best. I’ve rambled for many pages now and given lots of analogies and intuitions and models but nothing like a proof or decisive empirical study.

However, I think it’s on the right track. I also think that writing and thinking through this theory has helped me to better understand what various other people have said about agency (e.g. Yudkowsky/Veedrac), and perhaps reading it will have the same effect in others. I’d be interested to hear if so; otherwise I’ll assume not…

I might end up adding more posts to this sequence, e.g. “Here’s a criticism of the theory that seems plausible to me” or “Here’s an elegant restatement” or “here’s an important implication I hadn’t noticed before.” But for now this is the last one.

Thanks again to those who helped me with these ideas, especially Ramana Kumar.