Sharing Guidelines: Do whatever you want
Author: Neel Nanda
Status: Partially published, the rest abandoned, extremely out of date

Introduction

Short Version (2.5K words)

How to read this post
- Caveats

Addressing threat models
- Treacherous Turns
  - The case
  - The work
  - Sub-model: Inner Alignment
    - The Case
    - The work
- You get what you measure
  - The case
  - The work
- AI caused coordination failures

Agendas to build safe AGI
- Iterated Distillation and Amplification
- AI Safety via Debate
- Make AI Human-Compatible
  - The Agenda
  - The Work

Robustly Good Approaches
- The case
- The Work

Field-Building

Deconfusion Research

Crucial Considerations
- Timelines
- Prosaic AI Alignment
- Continuous vs Discontinuous Takeoff
- How hard is alignment?

Conclusion

An Introduction to the AI Alignment Landscape

[Alternate titles - thoughts appreciated!]
Disclaimer: I am currently an intern at DeepMind, but this post is entirely my personal views and speculations, and nothing to do with my employer or my work there.

Intended audience: People who understand why *you might think that AI Alignment is important, but want to understand what AI researchers actually* do *and why.*

Epistemic status: My best guess.

Epistemic effort: ~20 hours of work, running it by ~20 people

Thanks to [TODO:] [long list of names] for careful thoughts and comments, and making this post dramatically better and more coherent.

Introduction

[Alternate first sentences - thoughts appreciated!]

What needs to be done to make the development of AGI safer? This is the fundamental question of AI Alignment research, and there are many possible answers.

I've spent the past few months trying to get into AI Alignment work, and broadly found it pretty confusing to get my head around what's going on. Anecdotally, this is a common experience. The best way I've found of understanding the field is by understanding the different approaches to this question. In this post, I try to write up the most common schools of thought on this question, and break down the research that goes on according to which perspective it best fits.

There are already some excellent overviews of the field: I particularly like Paul Christiano's Breakdown and Rohin Shah's literature review and interview. The thing I'm trying to do differently here is focus on the *motivations* behind the work. AI Alignment work is challenging and confusing because it involves reasoning about future risks from a technology we haven't invented yet. Different researchers have a range of views on how to motivate their work, and this results in a wide range of work, from writing papers on decision theory to training large language models to summarise text. I find it easiest to understand this range of work by framing it as different ways to answer the same fundamental question.

My goal is for this post to be a good introductory resource for people who want to understand what Alignment researchers are actually doing today. I assume familiarity with a good introductory resource, eg Superintelligence, Human Compatible or Richard Ngo's AGI Safety from First Principles, and that readers have a sense for what the problem *is* and why you might care about it. I begin with an overview of the most prominent research motivations

and agendas. I then dig into each approach, and the work that stems from that view. I especially focus on the different **threat models** for how AGI leads to existential risk, and the different **agendas** for actually building safe AGI. In each section, I link to my favourite examples of work in each area, and the best places to read more. Finally, as another way to understand the high-level differences in research motivations, I discuss the different underlying beliefs about how AGI will go, which I'll refer to as **crucial considerations**.

Short Version (2.5K words)

[TODO: Insert links to each section in the text]
I broadly see there as being 5 main types of approach to Alignment research. I break this piece into five main sections analysing each approach.
Note: The space of Alignment research is quite messy, and it's hard to find a categorisation that carves reality at the joints. As such, lots of work will fit into multiple parts of my categorisation.

1. **Addressing threat models:** We keep a specific threat model in mind for how AGI causes an existential catastrophe, and focus our work on things that we expect will help address the threat model.
2. **Agendas to build safe AGI:** Let's make specific plans for how to actually *build* safe AGI, and then try to test, implement, and understand the limitations of these plans. With an emphasis on *understanding* how to build AGI safely, rather than trying to do it as fast as possible.
3. **Robustly good approaches:** In the long-run AGI will clearly be important, but we're highly uncertain about how we'll get there and what, exactly, could go wrong. So let's do work that seems good in many possible scenarios, and doesn't rely on having a specific story in mind. Interpretability work is a good example of this.
4. **De-confusion:** Reasoning about how to align AGI involves reasoning about complex concepts, such as intelligence, alignment and values, and we're pretty confused about what these even mean. This means any work we do now is plausibly not helpful and definitely not reliable. As such, our priority should be to do some conceptual work on how to think about these concepts and what we're aiming for, and trying to become less confused.
    a. I consider the process of coming up with each of the research motivations outlined in this post to be examples of good de-confusion work
5. **Field-building:** One of the biggest factors in how much Alignment work gets done is *how many* researchers are working on it, so a major priority is building the field. This is especially valuable if you think we're confused about what work needs to be done now, but will *eventually* have a clearer idea once we're within a few years of AGI. When this happens, we want a large community of capable, influential and thoughtful people doing Alignment work.
    a. This is less relevant to technical work than the previous sections. I include it because I both think that technical researchers are often best placed to do outreach and grow the field, and because an excellent way to grow the field is by doing high-quality work that other researchers are excited to build upon.

Within this framework, I find the **addressing threat models** and **agendas to build safe AGI** sections the most interesting and think they contain the most diversity of views, so I expand these into several specific models and agendas.

Addressing threat models: There are a range of different threat models. Within this section, I focus on three threat models that I consider most prominent, and which most current research addresses.

1. **Treacherous turns:** We create an AGI that is pursuing large-scale end goals that differ from ours. This results in convergent instrumental goals: the agent is incentivised to do things such as preserve itself and gain power, because these will help it achieve its end goals. In particular, this incentivises the AGI to deceive us into thinking it is aligned until it has enough power to decisively take over and pursue its own arbitrary end goals, known as a **treacherous turn**. This is the classic case outlined in Nick Bostrom's Superintelligence, and Eliezer Yudkowsky's early writing**.**
    a. **Sub-Threat model: Inner Misalignment.** A particularly compelling way this could happen is **inner misalignment** - the system is itself pursuing a goal, which may not have been the goal that we gave it.. This is notoriously confusing, so I'll spend more time explaining this concept than the others. See Rob Miles' video for a more in-depth summary.
        i. **A motivating analogy:** Evolution is an optimization process that produced humans, but from the perspective of evolution, humans are misaligned. Evolution is an optimization process which selects for organisms that are good at reproducing themselves. This produced humans, who were themselves optimizers pursuing goals such as food, status, and pleasure. In the ancestral environment pursuing these goals meant humans were good at reproducing, but in the modern world these goals do not optimize for reproduction, eg we use birth control.
        ii. The core problem is that evolution was optimizing organisms for the objective of 'how well do they survive and reproduce', but was selecting them according to their *performance* in the ancestral environment. Reproduction is a hard problem, so it eventually produced organisms that were themselves optimizers pursuing goals. But because these goals just needed to lead to reproduction *in the ancestral environment*, these goals didn't need to be the same as evolution's objective. And now humans are in a different environment, the difference is clear, and this is an alignment failure
        iii. Analogously, we train neural networks with an objective in mind, but just select them according to their *performance* on the training data. For a sufficiently hard problem, the resulting network may be an optimizer pursuing a goal, but all we know is that the network's goal has good performance *on the training data*, according to our goal. We have no guarantee that the network's goal is the objective we had in mind, and so cannot resolve treacherous turns by setting the right training objective. The problem of aligning the network's goal with the training objective is the **inner alignment problem**.

2. **You get what you measure:** The case given by Paul Christiano in What Failure Looks Like (Part 1):
    a. To train current AI systems we need to give them simple and easy-to-measure reward functions. So, to achieve complex tasks, such as winning a video game, we often need to give them simple proxies, such as optimising score (which can go wrong…)
    b. Extrapolating into the future, as AI systems become increasingly influential and are trained to solve complex tasks in the real world, we will need to give them easy-to-measure proxies to optimize. Something analogous to, in order to maximise human prosperity, telling them to optimize GDP
    c. By definition, these proxies will not capture everything we value and will need to be adjusted over time. But in the long-run they may be locked-in, as AI systems become increasingly influential and an indispensable part of the global economic system. An example of partial lock-in is climate change, though the hidden costs of fossil fuels are now clear, they're so ubiquitous and influential that society is struggling to transition away from them.
    d. The phenomenon of 'you get what you measure' is already common today, but may be much more concerning for AGI for a range of reasons. For example: AI systems are a human incomprehensible black box, meaning it's hard to notice problems with how they understand their proxies; and AI capabilities may progress very rapidly, making it far harder to regulate the systems, notice problems, or adjust the proxies
3. **AI Influenced Coordination Failures:** The case put forward by Andrew Critch, eg in What multipolar failure looks like. Many players get AGI around the same time. They now need to coordinate and cooperate with each other and the AGIs, but coordination is an *extremely* hard problem. We currently deal with this with a range of existing international norms and institutions, but a world with AGI will be sufficiently different that many of these will no longer apply, and we will leave our current stable equilibrium. This is such a different and complex world that things go wrong, and humans are caught in the cross-fire.
    a. This is of relevance to technical researchers because there is research that may make cooperation in a world with many AGIs easier, eg interpretability work.
    b. Further, the alignment problem is mostly conceived of as ensuring AGI will cooperate with its operator, rather than ensuring a world with many operators and AGIs can all cooperate with each other; a big conceptual shift

Note that this decomposition is entirely my personal take, and one I find useful for understanding existing research. For an alternate perspective and decomposition, see this recent survey of AI researcher threat models. They asked about five threat models (only half of which I cover here), and found that while opinions were often polarised, on average, the five models were rated as equally plausible.

Agendas to build safe AGI: There are a range of agendas proposed for how we might build safe AGI, though note that each agenda is far from a complete and concrete plan. I think of them more as a series of confusions to explore and assumptions to test, with the eventual goal of making a concrete plan. I focus on three agendas I consider most prominent - see Evan Hubinger's Overview of 11 proposals for building safe advanced AI for more.

1. **Iterated Distillation and Amplification (IDA):** We start with a weak system, and repeatedly **amplify** it to a more capable but expensive to run system, and **distill** that amplified version down to one that's cheaper to run.
    a. This is a notoriously hard idea to explain well, so I spend more words on it than most other sections. Feel free to skip if you're already familiar.
    b. **Motivation 1:** We distinguish between narrow learning, where a system learns how to take certain actions, eg imitating another system, and ambitious learning, where a system is given a goal but may take arbitrary actions to achieve that goal. Narrow learning seems much easier to align because it won't give us surprising ways to achieve a goal, but this also inherently limits the capabilities of our system. Can we achieve arbitrary capabilities only with narrow techniques?
    c. **Motivation 2:** If a system is less capable than humans, we may be able to look at what it's doing and understand it, and verify whether it is aligned. But it is much harder to scale this oversight to systems far more capable than us, as we lose the ability to understand what they're doing. How can we verify the alignment of systems far more capable than us?
    d. The core idea of IDA:
        i. We want to build a system to perform a task, eg being a superhuman personal assistant.
        ii. We start with some baseline below human level, which we can ensure is aligned, eg imitating a human personal assistant.
        iii. We then **Amplify** this baseline, meaning we make a system that's more expensive to run, but more capable. Eg, we give a human personal assistant many copies of this baseline, and the human can break tasks down into subtasks, and use copies of the system to solve them. Crucially in this example, as we have amplified the baseline by just making copies and giving them to a human, we should expect this to remain aligned.
        iv. We then **Distill** this amplified system, using a narrow technique to compress it down to a system that's cheaper to run, though may not be as capable. Eg, we train a system to imitate the amplified baseline. As we are using a narrow technique, we expect this distilled system to be easy to align. And as the amplified baseline is *more* capable than the distilled system, we can use that to help ensure alignment, achieving scalable oversight.
        v. We repeatedly amplify then distill. Each time we amplify, our capabilities increase, each time we distill they decrease, but overall they improve - we take two steps forward, then one step back. This means that by repeatedly applying narrow techniques, we could be able to achieve far higher capabilities.
    e. **Caveat:** The idea I've described is a fairly specific form of IDA. The term is sometimes used to vaguely describe a large family of approaches that recursively break down a complex problem, using some analogue of Amplification and Distillation, and which ensure alignment at each step.
2. **AI Safety via Debate**: Our goal is to produce AI systems that will truthfully answer questions. To do this, we need to reward the system when it says true things during training. This is hard, because if the system is much smarter than us, we cannot

distinguish between true answers and sophisticated deception. AI Safety via Debate solves this problem by having two AI systems debate each other, with a third (possibly human) system judging the debate. Assuming that the two debaters are evenly matched, and assuming that it is easier to argue for true propositions than false ones, we can expect the winning system to give us the true answer, even if both debaters are far more capable than the judge.

3. **Solving Assistance Games:** This is Stuart Russell's agenda, which argues for a perspective shift in AI towards a more human-centric approach.
    a. This views the fundamental problem of alignment as learning human values. These values are in the mind of the human operator, and need to be loaded into the agent. So the key thing to focus on is how the operator and agent interact during training.
    b. In the current paradigm, the only interaction is the operator writing a reward function to capture their values. This is an incredibly limited approach, and the field needs a perspective shift to have training processes with much more human-agent interaction. Russell calls these new training processes **assistance games**.
    c. Russell argues for a paradigm with 3 key features: we judge systems according to how well they optimise *our* goals, the systems are uncertain about what these goals are, and these are inferred from our behaviour.
    d. The focus is on changing the perspective and ways of thinking in the field, rather than on specific technical details, but Russell has also worked on some specific implementations of these ideas, such as Cooperative Inverse Reinforcement Learning

Robustly good approaches: Rather than the careful sequence of logical thought underlying the two above categories, robustly good approaches are backed more by a deep and robust-feeling intuition. They are the cluster thinking to the earlier motivation's sequence thinking. This means that the motivations tend to be less rigorous and harder to clearly analyse, but are less vulnerable to identifying a single weak point in a crucial underlying belief. Instead there are *lots* of rough arguments all pointing in the direction of the area being useful. Often multiple researchers may agree on how to push forwards on these approaches, while having wildly different motivations. I focus on the 3 key areas of interpretability, robustness and forecasting.

Note that robustly good does *not* mean that 'there is no way this agenda is unhelpful', it's just a rough heuristic that there are lots of arguments for the approach being *net* good. It's entirely possible that the downsides in fact outweigh the upsides.

*(**Conflict of interest**: Note that I recently started work on interpretability under Chris Olah, and many of the researchers behind scaling laws are now at Anthropic. I formed the views in this section before I started work there, and they entirely represent my personal opinion not those of my employer or colleagues)*

1. **Interpretability**: The key idea of interpretability is to demystify the black box of a neural network and better understand what's going on inside. This often rests on the implicit assumption that a network *can* be understood. I focus on mechanistic interpretability, which focuses on finding the right tools and conceptual frameworks to interpret a network's parameters.

   a. I consider Chris Olah's Circuits Agenda to be one of the most ambitious and exciting efforts here. It seeks to break a network down into understandable pieces, connected together via human-comprehensible algorithms implemented by the parameters. This has produced insights such as neurons in image networks often encoding comprehensible features, or reverse engineering the network's parameters to extract the algorithm used to detect curves.
   b. The key intuition for why to care about this is that many risks are downstream of us not fully understanding the capabilities and limitations of our systems, and this leading to unwise and hasty deployment. Particular reasons I find striking:
      i. This may allow a line of attack on inner alignment - training a network is essentially searching for parameters with good performance. If many sets of parameters have good performance, then the only way to notice subtler differences is via interpretability
      ii. Understanding systems better may allow better coordination and cooperation between actors with different powerful AIs
      iii. It may allow a saving throw to detect misalignment before deploying a dangerous system in the real world
      iv. We may better understand concrete examples of misaligned systems gaining insight to be used to align them and better understand the problem.
   c. This case is laid out more fully in Chris Olah's Views on AGI Safety.
2. **Robustness**: The study of systems that generalise nicely off of the distribution of data it was trained on without catastrophically breaking. Adversarial examples are a classic example of this - where image networks detect subtle textures of an image that are imperceptible to a human. By changing these subtle textures, networks become highly confident that an image is eg a gibbon, while a human thinks it looks like a panda. More generally, robustness is a large subfield of modern Machine Learning, focused on questions of ensuring systems fail gracefully on unfamiliar data, can give appropriate confidences and uncertainties on difficult data points, are robust to adversaries, etc.
   a. Why care? Fundamentally, many concerns about AI misalignment are forms of **accident risk**. The operators are not malicious, so if a disaster happens it is likely because the system did well in training but acted unexpectedly badly when deployed in the real world. The operators aren't *trying* to cause extinction! The real world is a different distribution of data than training data, and so this is fundamentally a failure of generalisation. And better understanding these failures seems valuable.
      i. Eg, deception during training that is stopped once the AI is no longer under our control is an example of (very) poor generalisation
      ii. Eg, systemic risks such as all self-driving cars in a city failing all at once
      iii. Eg, systems failing to sensibly during unprecedented world events, eg a self-driving car not coping with snow in Texas, or a personal assistant AI scheduling in-person appointments during a pandemic
   b. Dan Hendrycks makes the case for the importance of robustness (and other subfields of ML)

3. **Forecasting**: A key question when thinking about AI Alignment is timelines - how long until we produce human-level AI? If the answer is, say, over 50 years, the problem is *far* less urgent and high-priority than if it's 20. On a more granular level, with forecasting we might seek to understand what capabilities to expect when, which approaches might scale to AGI and which will hit a wall, which capabilities will see continuous growth vs a discontinuous jump, etc.
    a. In my opinion, some of the most exciting work here is scaling laws, which take a high-energy physics style approach to systematically studying large language models. These have found that scale is a *major* driver in model performance, and further that this follows smooth and predictable laws, as we might expect from a natural process in physics.
        i. The loss can be smoothly extrapolated from our current models, and seems to be driven by power laws in the available data, compute and model size
        ii. These extrapolations have been confirmed by later models such as GPT-3, and so have made genuine predictions rather than overfitting to existing data.
        iii. Ajeya Cotra has extended this research to estimate timelines until our models scale to the capabilities of the human brain.
    b. The case for this is fairly simple - if we better understand how long we have until AGI and what the path there might look like, we are *far* better placed to tackle the ambitious task of doing useful work now to influence a future technology.
        i. This may be decision relevant, eg a 10 year plan to go into academia and become a professor makes far more sense with long timelines, while doing directly useful work in industry now may make more sense with short timelines
        ii. If we understand which methods will and will not scale to AGI, we may better prioritise our efforts towards aligning the most relevant current systems.
    c. Jacob Steinhardt gives a longer case for the importance of forecasting

Key considerations: The point of this post is to help you gain traction on what different alignment researchers are doing and what they believe. Beyond focusing on research motivations, another way I've found valuable to get insight is to focus on **key considerations** - underlying beliefs about AI that often generate the high-level differences in motivation and agendas. So in the sixth and final section I focus on these. There are many possible crucial considerations, but I discuss four that seem to be the biggest generators of action-relevant disagreement:

1. **Timelines:** How long will it be until we have AGI? Work such as de-confusion and field-building look much better on longer timelines, empirical work may look better on shorter timelines, and if your timelines are long enough you probably don't prioritise AI Alignment work at all.
2. **Prosaic AI Alignment:** To build AGI, we will need to have a bunch of key insights. But perhaps we have *already* found all the key insights. If so, AGI will likely look like our current systems but better and more powerful. And we should focus on aligning our current most powerful systems and other empirical work. Alternately, maybe

we're missing some fundamental insights and paradigm shifts. If so, we should focus more on robustly good approaches, field-building, conceptual work, etc.

3. **Sharpness of takeoff:** Will the capabilities of our systems smoothly increase between now and AGI? Or will we be taken by surprise, by a discontinuous jump in capabilities? The alignment problem seems much harder in the second case, and we are much less likely to get warning shots - examples of major alignment failures in systems that are too weak to successfully cause a catastrophe
4. **How hard is alignment?:** How difficult is it going to be to align AGI? Is there a good chance that we're safe by default, and just need to make that more robustly likely? Or are most outcomes pretty terrible, and we likely need to slow down and radically rethink our approaches?

How to read this post

This post has ended up very long, so I've designed it to be skimmable and modular, rather than something you need to read fully or in order. I recommend reading the overview section, and then diving into the sections that feel least familiar or most confusing to you.

There's a lot of content written about alignment, and it can be hard to navigate and find the best work. In each section, I've collected links to my favourite examples of good work, sources that helped clarify my thinking, and places to read more. I think technical writing is often hard to digest, especially without a big picture in mind, so where possible I link to Alignment Newsletter summaries or Rob Miles videos for each piece of work. There's a lot of links in this post, so I recommend reading the summaries for anything that interests you, but being selective about what full-length works you read.

When reading a piece like this, it is easy to be passive rather than actively engaging with and questioning the ideas. I recommend noting down particularly interesting or confusing ideas as you go, and noticing what questions you feel uncertain or curious about. I cover many different perspectives on alignment in this piece, and a good end goal is to understand each one and why somebody might hold it, and to evaluate which ones you personally find more less compelling.

Caveats

- I'm pretty new to this field, and definitely misunderstanding a bunch of things. I expect I'll think fairly differently a few months from now! I would love to hear other people's opinions, and what you think I'm wrong about.
    - In particular, I try to summarise the views of many researchers in this piece. Summarising people's views is hard, and I don't claim these people would endorse my summary.
- There is a lot of disagreement about what "intelligence" or AGI even means. For simplicity, I will use AGI as a catch-all term for 'the kind of powerful AI that we care about'.
- Since I want this to be a good introductory resource, I should also emphasise what this post is *not*:

- I assume that readers are familiar with the basic ideas of AI Alignment, and are already convinced that existential risk from AGI is plausible. For a general introduction, I recommend Kelsey Piper's excellent explainer, and for an in-depth case for being concerned about AGI, I recommend Richard Ngo's excellent sequence
- This post is about different technical research directions for how to make AGI go well and does not try to cover **policy** and **governance considerations**. I think there are a *lot* of important considerations here for technical researchers, but I know embarrassingly little about this area. I recommend Allan Dafoe's overview and the Centre for the Governance of AI's Research Agenda (these cover both near-term and long-term considerations for making powerful AI go well)
- This post does not try to give concrete **career advice**: how to skill up, which organisations to apply to for which work, whether to do a PhD, etc. For these questions, I recommend Adam Gleave's guide to Beneficial AI Grad School, 80,000 Hours' guide to becoming an ML Engineer and Rohin Shah's FAQ. Skimming Larks' comprehensive annual literature reviews is a good way to get a sense for what different orgs are up to.

Addressing threat models

A common approach is to be specific, and focus on a threat model. To extrapolate from current work in AI and our theoretical understanding of what to expect, to come up with specific stories for how AGI could cause an existential catastrophe. And then to identify specific problems that make these failure modes more likely to happen, and try to solve it now.

It is obviously really hard to reason about the future in a specific way without being wildly off! But I am pretty excited about approaches like this. I think it's easy for research (or anything, really) to be meandering, undirected and not very useful, especially for vague and ungrounded problems such as AI Alignment. And having a specific story to guide what you do can be a valuable source of direction, even if ultimately you know it will be flawed in many ways. Nate Soares makes this case well.

Note that I think there is very much a spectrum between this category and robustly good approaches[TODO:link]. Most robustly good ways to help also address specific threat models, and many ways to address specific threat models feel robustly useful. But I find this a helpful distinction to keep in mind.

Treacherous Turns

This is the classic case outlined by earlier proponents of AI Alignment, especially Nick Bostrom and Eliezer Yudkowsky. It is outlined most clearly in *Superintelligence*. Joseph Carlsmith recently wrote a more up-to-date report examining a similar case.

The case

We produce an AGI. We believe this will be a goal-directed agent, trying to maximise a goal. Our current techniques cannot shape the goals of AIs very precisely, and worse, human values are highly complex and nuanced and vary between people, making them extremely hard to specify precisely. Further, maximising most large-scale goals means the AGI will have many instrumentally convergent goals - it will want to gain power, influence, resources and avoid being turned off, because these are instrumentally helpful for a wide range of tasks.

As goal specification is so hard, the AGI will inevitably want different things from us. It will have superhuman planning capabilities, meaning it will be better at coming up with ways to get what it wants than we will. And so it will likely come up with creative plans that we cannot predict and successfully guard against, because it is very hard to outwit something significantly smarter than you. A specific way this could go wrong is by creating an inventive to deceive us, to act perfectly aligned and to pass all tests we give it, until it can gain enough influence to decisively take power: a **treacherous turn**. This is not necessarily how things would *actually* go down, the key point is that if a system is better at planning than us, has different goals, and can influence the world, this can go wrong in many catastrophic ways.

I overall find this case fairly persuasive, and I expect there are significant grains of truth in this. It is by far the oldest and most established of the threat models I discuss, and has seen far more rigorous treatment than the others, but could still do with significantly more study. In particular, simplistic discussions of this model often bake in significant implicit assumptions, and it has often faced criticism.

Some criticisms:

- This framework tends to think of intelligence as a mysterious black box that caches out as 'better able to achieve plans than us', without much concrete detail. Further, it assumes that *all* goals would lead to these issues.
    - The sections on Goals and Agency in Richard Ngo's AGI Safety from First Principles do a good job of disentangling this.
- This case has been around since well before deep learning came on the scene, and some implicit ideas in earlier versions of the arguments now seem less plausible:
    - ‘Expected utility maximiser’ does not seem to describe modern systems (or humans!) very well.
    - It was previously believed that systems, when they reached human level, would learn to edit their source code. And thus make themselves smarter, become better at editing their source code, etc, leading a rapid rise in capabilities from human level to vastly superhuman: an **intelligence explosion**. ML systems are different as they need to first be trained, which takes a lot of time and compute, making it less likely that there could be such a big discontinuity in capabilities, and making it less likely that we get caught by surprise.
    - See Tom Adamczewski’s discussion of how arguments have shifted
- Ben Garfinkel has been a prominent critic of the public case for this model, and points out a range of other holes.

The work

- Understanding the incentives and goals of the agent, and how the training process can affect these in subtle ways
    - Work on specification gaming (aka reward hacking) - how AI finds ways to optimise reward functions in unexpected and perverse ways
    - Causal influence analysis work from Tom Everitt and Ryan Carey - using causal influence diagrams to better understand how subtle details of the training process can significantly affect the resulting incentives of the agent
- **Limited optimization:** Many of these problems inherently stem from having a goal-directed utility-maximiser, which will find creative ways to achieve these goals. Can we shift away from this paradigm?
    - **Satisficers**: Rather than making an optimizer striving to do *as well as possible*, make an agent trying to do 'good enough'
        - Jessica Taylor's Quantilizers are a cool formalisation of this - find the best policy that doesn't deviate too much from what a human would do
    - **Imitation**: Train agents to imitate humans. A human wouldn't try to take over the world, so an excellent imitator wouldn't.
        - See this post for discussion for and against
    - **Myopic agents**: Give an agent an inherently time-bounded goal, eg 'maximise reward over the next minute'. This time scale is too short for large scale planning, deception etc to make sense, but may still be useful for us.
        - See this post + comments for discussion for and against
- **Aligning AIXI:** AIXI is a theoretical ideal of a Bayesian reinforcement learning agent, and still has these problems of instrumentally convergent goals and power-seeking behaviour. So a theoretical angle of work is to try defining an aligned version of AIXI, and proving that this works. We can think of any RL system as an approximation to AIXI, and wouldn't expect an approximation to an unaligned ideal to be aligned itself, so solving this could be significant progress.
    - Michael Cohen does good work here.

Sub-Threat model: Inner Alignment

A particularly concerning subset of this problem is **inner misalignment**. This was an idea that had been floating around MIRI for a while, but was properly clarified by Evan Hubinger in Risks from Learned Optimization.

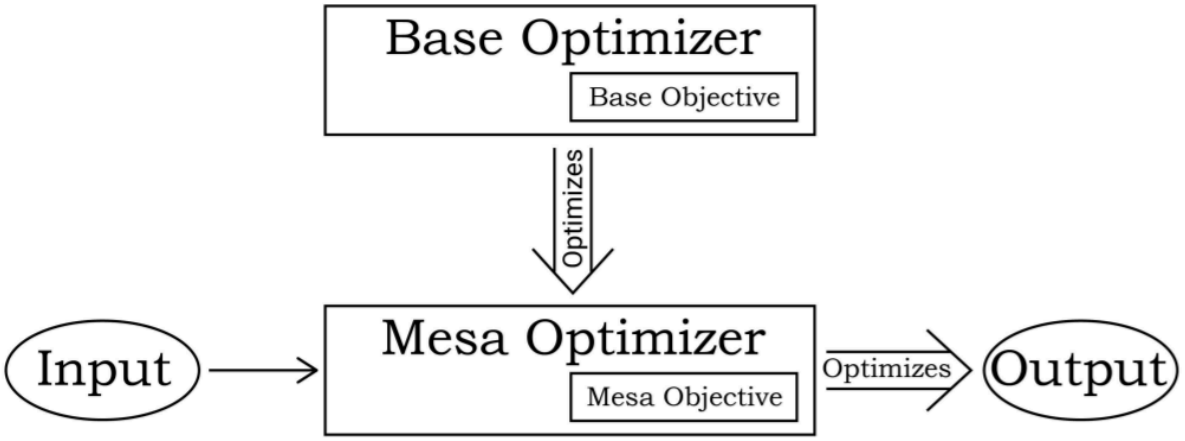
I think this is extremely important but hard to get your head around. Sources to learn more: Rob Miles' video, Evan's interviews on FLI and AXRP, the Risks from Learned Optimization paper.

The Case

We first begin with the analogy of humans and evolution: From a certain point of view, evolution is an optimization process that searches over the space of possible organisms and finds those that are good at reproducing. Evolution eventually produced humans, who are themselves optimizers, and we care about a range of goals, such as status, pleasure, art, knowledge, writing posts for the Alignment forum, etc. And in the ancestral environment,

pursuing these goals resulted in significant reproductive success. But in the modern world we continue to optimize *our* goals, yet totally fail to maximise reproductive success, eg by using birth control. Thus, from the perspective of evolution, humans are misaligned.

The key feature of the setup here, is that we had a **base optimizer**, evolution, an optimization process searching over possible *systems* according to how well they performed on a **base objective**, reproductive success. And this base optimizer eventually found a system, humans, that was itself optimizing. Humans are an example of a **mesa-optimizer**, an optimizing system found by a base optimizer, and humans are pursuing **mesa objective(s)**.



The core problem is that the base objective (reproductive success) and the mesa objective (status, pleasure, etc) are not the same. This happened because evolution only cares about the *performance* of a system *in the ancestral environment*, rather than what the system's mesa-objective truly is. And there are many possible mesa-objectives that will lead to reproductive success in the ancestral environment, but may lead to totally different outcomes in other environments - as happened with humans.

This setup is similar to modern deep learning: we search over possible neural networks weights with stochastic gradient descent (SGD), the base optimizer, according to our loss function, the base objective. And, further, SGD only pays attention to a network's *performance* on the loss function *on the training data*. And pays no attention to *how* a network actually works. So the concern is that deep learning may result in neural networks that are optimizing systems pursuing mesa-objectives, but have no way of ensuring these objectives are the same as the base objective.

This concern introduces significant further complexity into the alignment problem. Optimization is scary, and a highly capable system pursuing an objective misaligned with ours will likely lead to bad outcomes. But with mesa-optimizers, we have two objectives: the base objective, and the mesa-objective. So we need to both solve the **outer alignment problem**, ensuring the base objective is aligned with human values, and the **inner alignment problem**, ensuring that the mesa-objective is the same as the base objective.

A key feature of the inner alignment problem is that the base objective *underdetermines* the mesa-objective. Our main tool for reasoning about the outcome of training a neural network is evaluating which parameters lead to good performance on the training data. This tool breaks down here, as there are likely many mesa-objectives that perform well on the base objective on the training data, some of which will be aligned (as in, they generalise safely to new environments), some of which will not be. So the key question is which mesa-objective we will end up with.

In practice, if we end up with mesa-optimizers, they will have performed well on the base objective on the training data. There are many ways this could happen, here are three of the most important:

- **Inner alignment**: The mesa-objective and base objective are the same
- **Proxy alignment**: The mesa-objective is a good *proxy* for the base objective on the training data, but looks different in other environments. This is what happened with humans and evolution.
- **Deceptive alignment**: As an instrumentally convergent goal, the mesa-optimizer learns to deceive the base optimizer. It learns what the base objective is, and perfectly optimizes the base objective while being trained. Because the system performs well in training, it is eventually deployed, but once deployed it commits a treacherous turn.
    - This is a particularly scary outcome, because proxy alignment may be fairly easy to notice with sufficient testing, eg running the system in a wide range of simulated environments. Noticing deceptive alignment requires us to be able to trick the mesa-optimizer into *thinking* it is no longer in the training environment, which may be much harder.

I find this idea very compelling, but a lot of the discussion seems divided and confused, especially regarding what terms like optimizer actually mean. (See eg Evan Hubinger's Clarifying Inner Alignment terminology). And while humans fail inner alignment, humans do not seem like an expected utility maximiser. Personally, I'm not convinced that we will ever produce neural networks that act like expected utility maximisers. The main insight I've taken from inner alignment concerns is that modern deep learning only evaluates systems according to their performance on an objective, not their 'cognition': *how* they achieve that performance. There are many kinds of cognition that achieve good performance, and we don't know which a network is applying, making it much harder to see if it's truly aligned. And so I am very interested in research that may help us understand neural network cognition.

Another framing that side-steps the question of defining optimization is the 2D model of robustness. When we successfully train a model to act in an environment, it will take purposeful actions to achieve the intended objective. But when we shift the model to a different environment, there are three things that can happen. It may fail to take any purposeful actions at all, it may take purposeful actions but *not* towards the intended objective (its **capabilities** have generalised but its **objective** has not) or it may take purposeful actions towards the intended objective (its capabilities and objective have generalised). This breaks the question of 'does the model successfully generalise?' into the questions of 'does the model's **capabilities** generalise?' and 'does the model's **objective** generalise?'. This is a helpful distinction, because causing an existential catastrophe is really

hard, and so is much more likely to occur from an agent taking purposeful actions and capable of planning.

The work

This is a new and fairly poorly understood problem - it's not even obvious that we *will* get mesa-optimizers - so I divide the work into understanding the problem and solving the problem.

Understanding:

- Better understanding how and when mesa-optimization arises (if it does at all).
    - Eg, researching which training processes make mesa-optimization more or less likely to occur.
- **Empirical data:** Actually making concrete examples of mesa-optimizers
    - It's very hard to tell whether a neural network is actually an optimizer, so the main currently tractable approach is gathering empirical data of models whose capabilities generalise but whose objectives don't [todo: Add link to summary]
- **Inductive biases:** A neural network is a parametrisation to a space of functions. There are many different functions that all fit the training data equally well, but perform differently outside of the training data. To understand what happens, a key problem is understanding the **inductive biases** - when we train a network, we know we'll end up with weights with good performance on the training data, but how does a network choose *which* weights with good performance to return? If some weights give mesa-optimizers and others don't, which will be output?
    - For discussion of under-specification and how this affects modern ML, see this paper
    - A key part of the puzzle is **deep double descent** - in classic statistics, having more parameters means worse test set performance because you overfit. In deep learning, this trend eventually starts to reverse, when you have *lots* of parameters performance starts to improve again.
        - My rough intuition is that when there are lots of parameters, there are many possible functions. The inductive biases of a neural network favour simpler functions, and simpler functions generalise better to the test set, so having more functions to choose from leads to simpler functions which generalise better. But this is poorly understood, and that intuition could be totally off.
    - An intuition I find compelling is thinking of neural networks as Bayesian - that they have a prior over possible network parameters, and update this prior based on how well network parameters fit the training data. Under this model, understanding inductive biases reduces to understanding the neural network prior. In reality, neural networks learn by stochastic gradient descent, not by Bayesian updating, but they can be roughly modelled as sampling their trained parameters from the posterior distribution over the space of possible parameters.
        - See Alignment Forum discussion of a paper examining this question - I highly recommend reading the comments
    - There are also theoretical angles here. Eg, does AIXI have mesa-optimizers? - intuitively, if we search over all possible computer programs for ones that

perform well on our loss function, will we end up with computer programs that are mesa-optimizers?

Solving:

- **Interpretability** looks particularly good here, anything to make models less black box, so we could look inside and see what the model is optimizing. This both lets us check whether there *is* a mesa-optimizer and whether its objective is aligned, and hopefully to develop and verify solutions to this problem.
    - I discuss actual work on interpretability in the section on robustly good approaches[todo: link]
    - One way you might try to solve this is with black box interpretability techniques, essentially creating a test set to catch bad behaviours. Eg simulating a situation where the model can't be stopped, and seeing if it commits a treacherous turn. Some argue that this approach is limited as there is *inevitable* distributional shift between the test set and the real world, and that a deceptively aligned mesa-optimizer could act perfectly aligned until seeing an unfakeable sign of being out of training, meaning we *need* better interpretability techniques.
        - One example of this from Paul Christiano: RSA-2048 is a known number that is the product of two extremely large but unknown prime numbers. We do not know the factorisation, and do not have the computing power to calculate it now, but likely will in the next 100 years. So 'have I seen a factorisation of RSA-2048?' is an example of unfakeable evidence that the model is deployed in the real world, and not in a simulation.
    - For more surface area on how interpretability tools could be helpful here, I recommend skimming Evan Hubinger's 11 Proposals for Safe Advanced AI - for each proposal, he discusses how inner alignment might be solved
- This is highly related to robustness, especially the problem of distributional shift. It's fundamentally about systems that behave differently in training and the real world
    - Tools like relaxed adversarial training could help here.
    - (Note: I expect there's a lot of important work here - this section is short because I know very little about robustness, not because there is no work to include)

You get what you measure

This is the threat model outlined in What Failure Looks Like (Part 1) by Paul Christiano. I found the post insightful, but also somewhat cryptic, and found these clarifications from Ben Pace and Sam Clarke helpful. Paul Christiano's Another (outer) alignment failure story is another story outlining a related threat model, which I also found helpful. There has been less work on this threat model than the ones above, so the following is more my interpretation and attempt to steelman the ideas, rather than just a summary.

The case

Reinforcement learning systems are great at optimizing simple reward functions in clever and creative ways, and getting better at optimizing all the time, but we struggle to optimize

complex reward functions, and are seeing much less progress there. As I systems become more influential on the world and a bigger part of the global economic system, we will want them to achieve complex and nuanced goals, as human values are complex and nuanced. Assuming that we are much better at achieving simple rewards, this means we will need to *approximate* our true goals and define a **proxy** goal for the system. And if enough optimisation power is applied to a proxy goal, eventually these imperfections will become magnified, resulting in potentially catastrophic outcomes.

This pattern of simple, easy-to-measure proxies to achieve complex goals is widespread. For example, GDP is often used as an easy-to-measure proxy for measuring prosperity - this can work fairly well, but misses out on major components such as life satisfaction. Or, academia is intended to be a system to produce good science and advance human knowledge, by incentivising academics to publish rapidly, get many citations and publish in high-impact journals - this one often fails.

The notion of simple vs complex reward functions is doing a lot of work here, and is hard to define explicitly. Intuitively, I think of simple as “easy-to-measure” - could I give a system lots of samples from this reward function while training? In practice, reinforcement learning systems are often trained from very easy to measure functions, such as the score in a video game. It may be *possible* to train a system on more complex rewards, eg by having it directly ask a human for feedback, but we need systems trained on these complex rewards to also be *competitive* with systems trained on simple rewards - can we get comparable performance at comparable cost?

This phenomenon is not specific to AI, the world is already heavily shaped by systems optimising simple proxies, eg corporations maximising profit, and this is not (yet) an existential catastrophe. So why be concerned about AI?

One major reason for this is not currently a catastrophe is that society shapes and updates these proxies as the imperfections become clear through tools such as regulation. For example, ‘maximise profit’ is a bad proxy for ‘make society better’ as it doesn’t account for costs to third parties such as pollution, which can be addressed by taxes. But this error-correction mechanism may break-down for AI. There are three key factors to analyse here: **pace**, **comprehensibility** and **lock-in**.

Pace: How rapidly is the technology being developed and deployed? When trying to react to and regulate new technologies, it is *much* harder when things are moving at a fast pace - when things are slow, you have more time to react, coordinate, learn from failures, etc. For example, governments are having a really hard time regulating new technologies like drones and social media. AI is developing extremely rapidly even today, and if it becomes a significant fraction of global GDP this could plausibly be much worse, as far more resources will be put into it. (Note: This is not an argument for discontinuous/fast takeoff, a ‘slow’ takeoff would still likely be very hard to respond to)

Comprehensibility: Can we see what the system is doing and why? If so, it’s much easier to identify problems and notice them early. For example, regulating recommender systems is particularly hard because it’s hard to tell how the algorithm is making decisions, eg concerns around the Facebook algorithm radicalising people. A related point is that when there is a

problem that will require coordination and decisive action to solve, this is *much* easier with legible, uncontroversial and early evidence. For example, smoking is terrible for you, but it took a long time to realise this and discourage use because the link to lung cancer is noisy and acts on long time horizons. AI is currently mostly an incomprehensible black box, and will likely remain that way without significant progress in interpretability.

Lock-in: Once we've noticed problems, how difficult will they be to fix, and how much resistance will there be? For example, despite the clear harms of CO2 emissions, fossil fuels are such an indispensable part of the economy that it's incredibly hard to get rid of them. A similar thing could happen if AI systems become an indispensable part of the economy, which seems pretty plausible given how incredibly useful human-level AI would be. As another example, imagine how hard it would be to ban social media, if we as a society decided that this was net bad for the world. See Sam Clarke's excellent post for more discussion of examples of lock-in.

So, how bad is all this? My personal take is that an inappropriate focus on optimising metrics is clearly already happening in the world today, is causing many bad effects (and many good ones!) and that AI will plausibly make this worse. But it is highly unclear that this actually results in existential risk. Maybe the AIs will cause terrible collateral damage to eg the atmosphere or drinkable water (see discussion), maybe they'll never cause a catastrophe but result in the lock-in of suboptimal values, maybe they'll cause a bunch of short-term damage but we'll manage to fix things. It's very unclear! This case hasn't been very well fleshed out, though some researchers I respect take this very seriously - it was narrowly rated the most plausible threat model in a recent survey.

The work

- One of the most promising directions I've seen to this is directly training systems to get feedback from human operators, and learn to optimise for that feedback good - this lets the reward signal be anything that humans can judge, and allows for much more complex rewards. This is the field of **deep reinforcement learning from human feedback**.
    - A core difficulty here is that humans are expensive, and ML systems need a lot of data, so we need to become more data-efficient. One approach is to create a **reward model** based on small amounts of human feedback, train the system by querying the model, and asking the human for feedback on the most uncertain data points. An important paper here is Deep RL from Human Preferences, which managed to use just 900 bits of human feedback to teach a (humanoid) noodle to backflip
    - The OpenAI Alignment team has continued to do great work here, with two papers on teaching language models to summarise text based on human feedback - it's very difficult to code a good reward function for 'is this text a good summary' but you know it when you see it, so this was a meaningful advance on what we can do with simple reward functions.
        - One interesting finding here was that that the quality of human feedback *really* matters - performance went up significantly when they worked more closely with contractors to explain the task, and paid by

the hour rather than by the summary (incentivising more thorough work)
        - Another interesting finding was that if the system tries to *optimise* for its reward model, it will overfit and produce garbage. But if we regularise by constraining it to not deviate too much from a known OK policy, and then optimise for the reward model, it does great.
    - A more recent and more directly useful advance is the OpenAI Instruct Series - versions of GPT-3 fine-tuned to be good at following instructions and being helpful (currently available on the Beta API)
- Jan Leike, who now runs the OpenAI Alignment team, has an agenda of **recursive reward modelling**: This is a more powerful technique than just reward modelling, and resolves the problem of modelling complex reward functions by recursively breaking the problem down into smaller pieces, training reward models for *each* of those, and combining them into a reward function for the original problem. This is a particularly exciting approach, because as capabilities advance we could use this to create better reward models and thus safer systems. (See his paper and interview for more details)
    - Arguably, recursive reward modelling is a fully-fledged agenda to create aligned AGI. I categorise this as a way to help with this specific threat model, because I expect better reward models to be helpful across many paths to AGI
- Ajeya Cotra recently wrote up an agenda for this kind of work on aligning narrowly superhuman AI. Cutting edge systems such as GPT-3 are interesting, because alignment (in the sense of 'getting the system to do what we want') is becoming a bottleneck on the tasks we can use the system for, rather than capabilities (in the sense of 'what does the system know *how* to do'). For example, GPT-3 has much better medical knowledge than most doctors, but lies all the time, so currently cannot be used to safely give medical advice - this is an alignment problem, not a capabilities problem. The agenda argues that we should take these systems, and work to get them using their existing capabilities to their fullest extent, and learn how to train the system to do fuzzy and hard-to-specify tasks. And that this will likely give us good feedback on which alignment techniques actually work, make current systems safer, and make discontinuities in capabilities less likely.
    - The main work I have seen on this is deep RL from human feedback, but I can imagine other directions being fruitful!
    - I feel excited about this research direction, and as there is much more short-term economic incentive for this work than most alignment researcher, I hope there will soon be good work from non-longtermist researchers on it (though see Ajeya's case for why this is not *that* economically incentivised)
- There's a bunch of great work from non-longtermists on this front, e.g. the field of explainable AI, and the field of fairness and algorithmic bias seem highly relevant, though I know less about them

AI caused coordination failures

This is the case I've seen most pushed by Andrew Critch and David Krueger. They discuss it in their ARCHES paper, and Critch discusses it on the FLI podcast and in What Multipolar Failure Looks Like. This is also something that Allan Dafoe works on, and it is discussed in Open Problems in Cooperative AI and worked on by the new Cooperative AI Foundation. There hasn't been that much work fleshing out this case, and I don't understand it as well as I'd like to, so the following is my interpretation and my best attempt to steelman the core ideas, rather than solely my attempt at a summary.

The Case

This threat model stems from seeing cooperation and coordination failures as a fundamental lens through which to understand the world. Cooperation is hard and unstable, and coordination failures are the default state of the world, yet successful cooperation is the root of much of the value in the world. The concern centres around AI destabilising the current institutions and norms that enable cooperation, and causing coordination failures. I see this less as a single coherent case and more as a general prior that cooperation is hard yet crucial, and that destablisation will be bad. There are a bunch of specific points and stories, but I think you can disagree with those while buying the overall case. Note: Cooperation here can encompass cooperation between humans, between AIs, and betweens humans and AIs

Some rough intuitions for a cooperation-centric lens: Cooperation is unstable, because this involves many actors working together, where each actor is self-interestedly incentivised to defect, in a way that causes costs to others. Enough actors are self-interested that you need good institutions and norms to avoid them defecting. And most of human history is defined by being in a perpetual state of war and conflict. In modern times, some coordination failures have been *extremely* bad, eg WW1 and WW2, climate change, air pollution, etc. While when we can get cooperation right, eg trade, peace, well functioning governments, etc, this is responsibility for a lot of the progress humanity has made.

So, why would AI make cooperation worse/harder?

- **Pace**: AI systems will likely develop rapidly in capabilities, giving us less time to get used to them. And they will likely be able to think much faster than humans. Much of the cooperation in the world today comes from norms and institutions that are slow to develop and take time to build, this is much harder in a rapidly moving world
- **Destabilisation**: A world with AI will be very different to the world today. This likely means that different people, countries and institutions will have power. We are currently in something like a stable and reasonably cooperative equilibrium, but because cooperation is really hard, there is no reason to expect that a world with AI will be as cooperative as today's world
    - In particular, if established power structures are overturned, then many entities *could* get a lot of power if they act decisively, which encourages reckless behaviour from many actors. Unlike today's world, with a small handful of established superpowers (China, USA, Russia, etc)
    - Another angle on this is if AI and technological progress changes the offence-defence balance of technology. Eg, if it turns out to be extremely easy to create tiny autonomous drones to assassinate people, and really hard to defend against this. Or if AI makes it much easier to create extremely

persuasive arguments and videos (a la deepfakes) and this is hard to defend against, this may create a breakdown of trust and public discourse
- Institutions that work well on humans may fail to transfer to AIs, eg it's unclear what successful law enforcement would look like on an autonomous system

Maybe you agree that cooperation would be harder, and that this would be bad. But would this lead to an existential risk? I personally find this fairly unclear, and don't feel very compelled by any particular story, but I find it plausible that this could lead to extremely bad outcomes. See Section 3 of ARCHES and What Multipolar Failure Looks Like for more discussion.

One significant risk centres around collateral damage, the side effects of the coordination failures cause damage to eg the atmosphere or drinkable water and this causes humanity to die out. The underlying intuition here is one of **human fragility** - there is a large range of possible ways the Earth could be (temperature, composition of the atmosphere, etc) that could lead to machines thriving, while humans need a fairly specific environment. This means that unless AIs make a special effort to keep the Earth a good place for human life, and care highly about this, this will likely be expensive to maintain, and we should not expect this to go well by default.

This feels related to the concerns outlined in 'You get what you measure'[todo: link to section once published], but different. Before, AI systems cause collateral damage because the damage was instrumentally useful to their goals, and they weren't programmed to care about the harms. Here, each agent may care about the harms, but not enough - if each agent only plays a small marginal part in the coordination failure, they may not be incentivised to change it. Eg, each agent may find it valuable to burn fossil fuels, and accepts their marginal contribution to climate change, even if all agents are opposed to the overall effects of climate change.

Another risk is that when multiple AI systems are interacting, their interactions can cause unexpected feedback loops that bring them far outside of their training distribution, resulting in extreme and unexpected behaviour. One mundane example of this is the 2010 Flash Crash where interactions between badly programmed stock market trading bots resulted in a crash, which wiped trillions of dollars of value in minutes before recovering. A more speculative version of this raised by Critch is the Production Web, where the entire economy becomes automated and dominated by AIs, which build on each other and cause a great deal of growth, but does not cache out as morally relevant things such as increased human welfare, and ceases to be human comprehensible. This seems particularly concerning with new ideas such as Decentralized Autonomous Organizations, which could result in an economy where most resources are not, ultimately, controlled by humans, which could become totally out of control.

A useful framework introduced in ARCHES for thinking about AI is that of a **delegation problem** - operators make AI systems, and want the AI systems to act to achieve their values. This delegation problem is very different if there is one or many operators, and one or many AIs, giving four different scenarios:

- **Single-single** - The classic conception of the alignment problem, where a single operator wants to align a single AI with their values. This is already extremely hard!

- **Single-multi** - A single operator is using many AI systems, and wants the resulting behaviour to achieve their values. This is a different problem from single-single alignment, because we know from game theory that a system of rational agents acting autonomously rarely maximises total value (eg the prisoner's dilemma), and this often needs better coordination mechanisms
- **Multi-single** - Many operators want to use a single AI system to achieve their values. Just figuring out what the AI system should be doing is difficult! (Value aggregation is a notoriously hard problem)
- **Multi-multi** - the problem of many operators, and many interacting AI systems, where each operator likely has many AIs of their own, and we want the overall system to result in good outcomes. This sounds extremely hard!

One insight from this framing is that we should expect the multi-multi delegation problem to be neglected. The Alignment problem, as normally conceived, is single-single delegation, and it is plausible that the creators of AGI will put significant resources into solving it (though likely not enough!), since it is clearly their responsibility. But no one has a clear responsibility to solve multi-multi delegation, and far fewer resources are invested into it today. Yet, given how much greater train-time compute is than run-time compute for modern ML, it seems likely that once we have AGI, we will run many copies of it, and plausible that many different actors will have access to AGI. This means that the multi-multi delegation problem will become relevant at almost exactly the same time as single-single, and is plausibly a much harder problem.

The Work

I expect much of the useful work here to be policy centred, eg creating good institutions, regulations and norms around AI use, incentivising international cooperation, etc. But the proponents also argue that there is important technical work to be done to create AI agents better able to cooperate, which is what I'll focus on here:

- A lot of existing mainstream ML research seems relevant to this. In ARCHES and an accompanying blog post Critch ranks different subfields of ML and safety according to their relevance to enabling cooperation. I was surprised by how highly he ranks fairness in ML, computational social choice theory and accountability in ML
    - Note: I expect the rankings in this post to be controversial, and are just one researcher's opinion, and you shouldn't update too much from just that post. I would give quite different numbers
- Open Problems in Cooperative AI is a recent paper trying to define a new research agenda of creating AI systems that better enable cooperation.
    - One notable thing about this paper is that most of the authors are, to my knowledge, DeepMind researchers who normally focus on Multi-Agent RL work, rather than safety.
    - This has since led to the Cooperative AI Foundation for encouraging and funding this research, and a Nature editorial
- An interesting research direction is formalising cooperation, what it means and how it should work. Critch is interviewed by Daniel Filan on AXRP about a related paper on negotiable RL

- The Centre for Longterm Risk works on similar issues around game theory and encouraging good bargaining and cooperation between AIs, as summarised in this post. Though they focus specifically on preventing failures of negotiation between AIs.
- I am fairly unfamiliar with this area, and have likely missed relevant things. The Alignment Newsletter summarises a lot of work to do with Handling Groups of Agents, and you can skim through that to get a sense for what other kinds of work happens

Agendas to build safe AGI

Another compelling, but rather more ambitious agenda is designing a specific scheme by which we might actually *build* a safe AGI. In many ways this is a far harder problem, but we may be more confident that this will actually result in safe AGI. Note, these are better thought of as research agendas, rather than concrete plans - there are many holes, uncertain assumptions, and room to fill in more details (obviously we don't actually know how to make an AGI yet!). I think of them as research agendas where the *ideal end goal* is a concrete plan to build safe AGI.

I think there is a fairly blurry line between this approach and the approaches discussed in addressing threat models [TODO: link], and it's worth contrasting the two. Both approaches are often about trying to solve a particular threat model (rather than an agenda that will solve *all* of the safety problems). I see the main difference as how specific a view you take about how we will get to AGI. Agenda-building approaches plan out a specific path to AGI that they hope will resolve the threat model, while approaches in the previous section are more about developing specific patches or techniques that could be applied across many different paths to AGI. Both of these approaches can be very valuable, and I want to see both pursued. Agenda-building is very ambitious and it is hard to identify a specific path to AGI that will work, but would be extremely impactful if it works, will need to be done eventually, and even if the agenda is not ultimately used, having concrete approaches in mind makes it much easier to identify and reason about safety problems. On the other hand, AGI is a distant future technology, meaning we should be highly uncertain about how we'll get there, and so it is valuable to do work that will be useful across many paths to AGI.

It's also worth emphasising that work can be both theoretical and empirical under both approaches. Addressing threat models can be about the theoretical work of e.g. reasoning about stories about risk, or it can be about the empirical work of e.g. developing concrete examples of these problems in real systems and devising and implementing techniques to fix them. Similarly, agenda building can be about the theoretical work of e.g. designing an ambitious new scheme that may result in alignment, or the empirical work of e.g. implementing this agenda in the best systems we have today, and testing the capabilities and limitations of these systems.

The three proposals I discuss here are just the three I know the most about, have seen the most work on and, in my subjective judgement, the ones it is most worth newcomers to the field learning about.

As such, this should not be taken as a comprehensive overview of *all* current proposals. I highly recommend Evan Hubinger's post 11 Proposals for Safe Advanced AI for giving a good overview of several more proposals. I particularly like how he breaks down four key properties of a proposal:

- Outer alignment - if this scheme trained the *optimal* system for this loss function, would that be aligned?
- Inner alignment - are we confident that this scheme won't lead to misaligned mesa-optimizers? [todo: Link to inner alignment section]
- Performance competitiveness - will this system compete in practice with less safe approaches?
- Training competitiveness - will it put you at a significant disadvantage to train this system, rather than less safe alternatives?

To me, one of the key properties of a scheme for aligning *super*intelligent AI is how we ensure that as the AI goes beyond human intelligence the system remains aligned, and I'll try to highlight where this comes from in each:

Iterated Distillation and Amplification

Iterated Amplification and Distillation (IDA) is Paul Christiano's brainchild, and I consider it one of the most compelling proposals thus far regarding how we might actually align AGI. Note that it is notoriously difficult to get your head around, so expect this section to be harder than most - feel free to skip!

My favourite introduction is Ajeya Cotra's, see also a video by Rob Miles and a (beautifully illustrated!) post from Rafael Harth. For a more in-depth explanation, see a full sequence from Paul Christiano. Alex Zhu's FAQ and Chi Nguyen's post may be useful for more in-depth clarifications.

To explain the key ideas, I'll copy my description from the short version:

- We start with a weak system, and repeatedly **amplify** it to a more capable but expensive to run system, and **distill** that amplified version down to one that's cheaper to run.
    - **Motivation 1:** We distinguish between narrow learning, where a system learns how to take certain actions, eg imitating another system, and ambitious learning, where a system is given a goal but may take arbitrary actions to achieve that goal. Narrow learning seems much easier to align because it won't give us surprising ways to achieve a goal, but this also inherently limits the capabilities of our system. Can we achieve arbitrary capabilities only with narrow techniques?
    - **Motivation 2:** If a system is less capable than humans, we may be able to look at what it's doing and understand it, and verify whether it is aligned. But it is much harder to scale this oversight to systems far more capable than us, as we lose the ability to understand what they're doing. How can we verify the alignment of systems far more capable than us?
    - The core idea of IDA:

        - We want to build a system to perform a task, eg being a superhuman personal assistant.
        - We start with some baseline below human level, which we can ensure is aligned, eg imitating a human personal assistant.
        - We then **Amplify** this baseline, meaning we make a system that's more expensive to run, but more capable. Eg, we give a human personal assistant many copies of this baseline, and the human can break tasks down into subtasks, and use copies of the system to solve them. Crucially in this example, as we have amplified the baseline by just making copies and giving them to a human, we should expect this to remain aligned.
        - We then **Distill** this amplified system, using a narrow technique to compress it down to a system that's cheaper to run, though may not be as capable. Eg, we train a system to imitate the amplified baseline. As we are using a narrow technique, we expect this distilled system to be easy to align. And as the amplified baseline is *more* capable than the distilled system, we can use that to help ensure alignment, achieving scalable oversight.
        - We repeatedly amplify then distill. Each time we amplify, our capabilities increase, each time we distill they decrease, but overall they improve - we take two steps forward, then one step back. This means that by repeatedly applying narrow techniques, we could be able to achieve far higher capabilities.
    - **Caveat:** The idea I've described is a fairly specific form of IDA. The term is sometimes used to vaguely describe a large family of approaches that recursively break down a complex problem, using some analogue of Amplification and Distillation, and which ensure alignment at each step.

(End copied section)

Note that the examples given are a fairly specific form of IDA. The term can be used to describe a large family of approaches that may be applied to different tasks, and with different techniques to amplify or distill, giving different alignment properties. See Proposals 2, 3, 4, 10 and 11 of Evan Hubinger's 11 proposals for building safe advanced AI for different examples, and discussion of how we might ensure inner alignment.

A particularly exciting aspect of IDA is the argument that the system solves **outer alignment** - that is, the idealized, limiting system for this training setup and objective will be aligned. This is a desirable property, because the *actual* systems we get will be an approximation to this idealised version. If the idealised version is aligned, we just need to ensure that the approximation errors do not lose us alignment - though of course this is still very hard. While if the idealised version is not aligned, we somehow need to ensure that the approximation errors *gain* us alignment, which seems fairly intractable.

So, what is the limiting case of IDA? In the idealized case, we can imagine that the distillation steps create perfect copies, so we focus on amplifications. Initially, we start with a human thinking for a short amount of time. After amplifying once, this human is consulting many copies of themselves, *each* thinking for that short amount of time. After amplifying twice, they consult copies, each of whom then consult *sub*-copies, and each sub-copy is

thinking. In the limiting case, we get an **infinite tree of humans**, each layer consisting of many people, each consulting multiple people on the layer below. This infinite tree is called **HCH** (a recursive acronym standing for Humans Consulting HCH), and is notoriously hard to think about - see this post for more detail.

Is HCH aligned? The main argument in favour is that, if we expect each human to be aligned, we should expect the infinite tree to be aligned. Of course, this is far from obvious, and not something we fully understand - it is not clear to me whether a single human being aligned means a finite tree of interacting humans will be aligned, let alone an infinite one. And even the claim 'humans are aligned' is reasonable to push back on.

To reiterate, outer alignment is a *theoretical* property. We do not actually expect to get HCH, but understanding the idealized case of our methods may be valuable for understanding what will actually happen. (Or may be too misleading to be useful! This stuff is hard)

Is HCH competitive? That is, can HCH solve all the problems that an unaligned AGI, trained with less safe techniques, could solve? This is essentially the question of, can all ideas, no matter how complex, be broken down into enough (albeit maybe very many) small chunks of eg one person thinking for 5 minutes. This is an open question, called the **factored cognition hypothesis**. See this post for a quick introduction, and this post and this sequence for a deeper dive. Personally, my initial reaction to the hypothesis was that it was obviously false, but it's more subtle than it seems - there are lots of problems that can't be broken down into a *few* small chunks, but I find it hard to reason about *arbitrarily many* small chunks. And of course, it might be true that *HCH* can solve any problem, without IDA being competitive with less safe techniques *in practice* - it may be too expensive to train, need too much human labour, we may never be able to get close to HCH, etc.

And IDA does not need to fully solve the Alignment Problem to be useful, and Paul Christiano has said he doesn't expect it to be a fully solution on its own - it may solve some parts of the problem, and leave other parts that must be patched with different alignment techniques. One idea I've seen raised for a problem that IDA cannot solve is inaccessible information. I don't fully understand this myself, but you can read more here.

What work is currently being done?

Regarding empirical work, Ought used to be doing experiments exploring the factored cognition hypothesis, though I believe they no longer work on this. Recursive reward modelling (a research agenda from Jan Leike) and AI Safety via Debate (a research agenda from Geoffrey Irving) have both had empirical work done on them, as discussed elsewhere in this post [TODO: link]. I've often heard it claimed that these two are similar to IDA, in the sense that all are about breaking down knowledge recursively and, stronger, that if a problem prevents one technique from producing safe AGI, that problem will likely apply to the other techniques too. Personally, I don't fully understand this belief, and haven't seen a public defence of it, though there are examples of eg work on debate giving insight into problems with IDA

I'm otherwise not very familiar with the work here, especially theoretical work. I recommend looking through the Iterated Amplification section of the Alignment Newsletter Database to explore related work that is happening.

AI Safety via Debate

Debate is the brainchild of Geoffrey Irving. For more details see this summary, the blog post and paper and his interview on the FLI Podcast. For discussion of pros and cons, see Richard Ngo's Why I'm Excited About Debate

The core goal is to produce an Oracle. We do this by taking a question, and then having AI generate arguments for and against that question, and have this debate judged by a human to decide who won. The hope is that it is fundamentally easier to argue for the truth than a lie. And so, if we train both agents with the objective to just win the argument (ie without caring about truth), randomly assign each to a side, and they are equally matched, then the winner will be the one arguing for the truth.

Interestingly, this scheme was also inspired by AlphaZero: A debate can be viewed as a zero sum, two player game, where actions as making statements in response to the other player. We've seen a lot of progress in zero sum two player games, and this allows for strategies such as generating training data via self-play, so this may make Debate more practical.

This is exciting, because it is easier to evaluate an argument than it is to generate it in the first place. Thus, hopefully, we could get **scalable oversight** - the ability to use superhuman language agents to generate knowledge in a way that is human comprehensible and verifiable. In the framework of a two player game, the game is played in a vast tree of all possible debates. Generating knowledge looks like navigating this massive tree. By having superhuman agents engage in debate, we explore a single path in this tree, but can safely assume that this path is the only interesting one - each agent will choose the strongest argument at each step.

To me, the important and speculative key claim of Debate is that it is *easier* to argue for truth than falsehood. If we look at human debates, this seems pretty clearly false. Often the truth is complicated, and effective rhetorical technique wins out. A case for optimism is that many things are easier in this setting. We can ensure the judges are competent and careful, and possibly even train AI judges. We can use techniques like **cross-examination**, where we make multiple copies of an agent, and ask each copy different questions. This stops tricks like subtly changing the topic of debate over time - we can load an earlier version and ask it to clarify what it meant by the claim it just made.

Debate is also interesting, because we can make progress on it without powerful language models, by replacing any of the debaters and judge with humans, and do empirical work. Beth Barnes is the main person I know of working on this, and seems to have found a bunch of problems, and some solutions! (See update 1 and 2)

Make AI Human-Compatible

The Agenda

This is the model put forward by Stuart Russell (one of the world's most prominent CS professors) in his book Human Compatible, and some of his recent academic work. I found this newsletter valuable for understanding this perspective, and I highly recommend reading the entire thing. I also found Rohin Shah's Value Learning Sequence helpful.

Russell criticises the prevailing paradigm of Machine Learning, what he calls the **standard model**: that "Machines are intelligent to the extent that their actions can be expected to achieve their objectives.". Instead he proposes a new way of defining the end goal, that he describes as "Machines are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives".

The key thing argued for here is a perspective shift. Russell thinks that the way people in ML think about designing systems will predictably lead to bad outcomes, and further thinks that if the culture and paradigm of ML could be shifted to be more human centric in this way, that this will likely lead to better outcomes.

To put this paradigm into practice, he suggests 3 principles:

1. The machine's only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior.

I find the framing of Assistance Games valuable for understanding this view. The key perspective shift is to focus on the relationship and interactions between humans and AI systems. Human preferences and values are the core source of alignment, and so to understand whether systems will end up aligned, it's important to see how interactions with humans are baked into the training and deployment processes.

In a standard ML system, they're baked in exactly once - when we design the system architecture, loss function and training data. After this, we train the system and deploy it in the world to maximise whatever objective it has received. We may try to adjust things afterwards - correcting or turning off systems that are going wrong, etc, but this is fundamentally tacked on. A symptom of this fundamental problem is the Stop Button problem - any system capable of it will try to stop us turning it off, because this will prevent the system from achieving whatever its goal is. As Russell says "you can't fetch the coffee if you're dead".

The key insight of an Assistance Game is that *the fundamental issue* is that human feedback is not baked into the training process. And that if we instead found ways to train systems where human feedback is an inherent part of the process, we might make systems that *want* to help us turn them off. Principle 2 is particularly important here - a system will only want to be turned off if it is uncertain, and thinks its operator has information that the system does not.

To me, this work feels like a combination of field-building (trying to change the paradigm of the field), focusing on specific threat models (systems leading to perverse outcomes because of slightly under-specified rewards), and actively presenting an agenda to lead to safe AGI.

The Work

One of the key difficulties here is that a lot of energy is going towards the 'standard model', and much less towards this approach. So a key step will be making this perspective compelling, tractable, and something people want to work on. Something that can redirect the energies of the field in a more human compatible direction.

One particularly interesting attempt to operationalise these ideas is **Cooperative Inverse Reinforcement Learning** (CIRL). Inverse Reinforcement Learning (IRL) is an approach to RL where, rather than specifying a reward function explicitly, the system observes another agent, tries to infer the other agent's reward function from its actions (usually assuming some kind of rationality principles to connect behaviour and goals), and then optimises this inferred reward function. With CIRL, we train a system to infer a human's goals, but the twist is that we don't just infer them by observing the human's behaviour, we assume that the human *wants* us to understand their goals, and will actively help us - correct misunderstandings, giving the most useful information, etc.

One of the important points in all this is *how* we infer human values from behaviour, and this seems an important area for future work. This often requires some assumptions of rationality, which is problematic, because humans are biased, noisy, boundedly rational agents in many ways. Stuart Armstrong argues that since humans are irrational, possible human values are under-determined if we only observe our behaviour, and that approaches like Occam's Razor (take the simplest set of values that explain our behaviour) are insufficient.

As far as I'm aware, this approach is mostly focused on addressing outer alignment, rather than inner alignment concerns.

(Note: I think this agenda describes Stuart Russell's work, and the work of *some* other people at CHAI, but CHAI is a big lab and has people working on a range of things with a range of motivations, eg interpretability and adversarial robustness)

Robustly Good Approaches

The case

If we want to influence the trajectory of human civilisation for the better, we should find things that will be *really* big deals, and ensure they go well. If we look back on our era 1000 years from now, and think about what matters, creating autonomous entities that are smarter, think faster and are more capable than us seems like one of the most important things happening. So ensuring AI goes well is a major priority! I think Buck Shlegeris' My Personal Cruxes for AI Safety is a good example of this kind of reasoning.

One version of this is the **second species** argument. We can think of autonomous AGI as a separate intelligent species that we're creating. If there are two species, and one is significantly smarter than the other, the default outcome is that the preferences of the smarter species determines how things go in the longterm. This is sometimes called the **gorilla problem**: humans bear gorillas no particular ill will, but gorillas today only really survive because we want them to. This is not an outcome we want for humanity!

But, while this is a story for why this is obviously important, it's quite vague. And this makes sense! It's really hard to reason about how transformative future technologies will play out. And even if we can come up with compelling and specific threat models, we're likely going to be missing some other important threat models. As a result, we should focus still focus on specific problems, but favour ones that are *robustly* good in many possible paths for producing AGI, and will help humanity handle this well.

For an excellent and more in-depth take on robust reasons to be concerned, I highly recommend Richard Ngo's AGI Safety from First Principles. (I'm not sure either Richard or Buck would endorse the second conclusion of 'we're confused and should prioritise things that are robustly good' though)

The Work

- I think interpretability is the clearest example of this kind of work - having a better sense of what neural networks are doing and why seems clearly helpful in most possible worlds.
    - Within interpretability, by far the most exciting work I've seen is Chris Olah's Circuits work (Summary: 1 & 2)
    - I think the field of Explainable AI is also pretty great from this perspective, though I am not aware of much longtermist work in this area.
    - See Peter Hase and Own Shen's recent literature review of interpretability from an Alignment perspective
    - Chris Olah's Views on AGI Safety is a great case for some robust and some specific ways that interpretability could be helpful
- Robustness: What are the ways AI systems can go wrong, and what can we do about this?
    - Adversarial examples are the most obvious exaple of this.
    - I'm excited about work on red teaming - finding ways that a neural network is not robust or can be tricked into doing bad things - and adversarial training - automatically doing this during training, such that we can train networks that are more robust.
    - More generally, the problem of robustness to distributional shift (See p16 here) and safe generalisation seem extremely important. We cannot train our AGI on the real world, and it will inevitably encounter novel situations when deployed - safe generalisation is about ensuring this goes well.
- Forecasting AGI, and thinking about timelines. Understanding how long we will have to prepare, what we can expect along the way, etc
    - Most of the best work I've seen in this area comes from the Forecasting team, formerly at OpenAI. Their work seems heavily based on the **scaling**

hypothesis, that the key to AGI is taking our current models and making them bigger. They've had a lot of success with Scaling Laws, discovering that the performance of most ML systems improve in extremely predictable ways as they scale up, as a function of model size, compute and data.
        - Ajeya Cotra's excellent report on timelines is both a good example of forecasting work, and relied heavily on this
        - See this piece by Gwern on GPT-3 as significant evidence for the scaling hypothesis
    - AI Impacts also seems to do excellent work in this area.
- Limited AI: Ensuring that we can control a AGI, no matter how aligned or misaligned it is
    - Is it possible to safely put an agent smarter than us into a box, and constrain it from interacting with the world?
    - How can we get the AI to understand side effects, and minimise its impact on the world
        - Alex Turner and Victoria Krakovna have done good work here

Field-Building

Another, fairly different approach to all this is **field-building**. The case here is that we're confident AGI will be a big deal, and think that one of the biggest levers will be having capable researchers doing alignment research. This means that a *major* factor in things going well is *how many* good alignment researchers we have. And building the field is a big force multiplier - if you can get a researcher as capable as you into the field, that's equivalently valuable to your *entire research career* (with a bunch of caveats, since counterfactual impact is hard to reason about)

This is particularly compelling from a perspective of high uncertainty about how we'll get to AGI, and long timelines. This means we can probably do much better safety work in, say, the 5 years before getting to AGI. This means we want as many researchers as possible to be familiar with the field and capable of doing good work when we get to crunch time. Since we'll have a much better idea of what work to do later rather than now, field-building seems robustly good, and a way to do good work now

I'm including field-building in a post aimed at technical researchers because I think that this can often be the main motivation behind technical work, especially work aimed at an academic audience. Alignment research has historically had a major image problem as seeming overly weird and sci-fi-ish. By doing excellent work that people get excited about, showing that this work can be published in major conferences, writing agendas that give people ideas to hook onto, etc, this can help encourage other researchers to transition into doing safety work, and make it seem an important and respectable thing to work on. My impression is that this is a reasonably big factor of CHAI's theory of change.

Field-building work thus far seems to have been incredibly successful! Most major conferences now have an official safety category for papers, the amount of safety work produced seems to have dramatically increased over time, and this seems to be slowly

entering the academic Overton Window. And AI Safety seems *vastly* more mainstream now than it was, say, 5 years ago

Another exciting angle here is doing research that unlocks new ways for other researchers to contribute to alignment work, especially researchers outside of mainstream ML. I think this is ambitious, but that setting an agenda is *much* harder than working on an existing agenda, and much easier to publish on, etc. Some of my favourite examples:

- Creating benchmarks and baselines to test new safety ideas, eg DeepMind's AI Safety Gridworlds and OpenAI's Safety Gym (see second highlight)
    - I'd love to see safety benchmarks for large language models, eg testing for deception
- Chris Olah's Circuits work
- Ajeya Cotra's The case for aligning narrowly superhuman models
- Geoffrey Irving & Amanda Askell's AI Safety needs social scientists
- Stuart Russell's attempt to change the 'standard model' (as described above[TODO: Link to section])

I also think that there's also a lot of less technical field-building that technical researchers are unusually well-placed to do - especially centred around mentoring, public outreach, and getting other technical people on board. Some categories: (note that, other than mentoring junior people, I expect many of these to only really be accessible to more experienced researchers):

- Broad public outreach, conveying technical ideas behind alignment well:
    - Popular books, such as Superintelligence, Human Compatible and the Alignment Problem seem great here, along with the authors generally being public intellectuals, doing interviews, etc.
        - Superintelligence, in particular, seems to have played a *massive* role in Safety becoming more mainstream
        - Anecdotally, I know multiple safety researchers who decided to get into the field after reading one of these books, or after being primed by one of these books and seeing a major advance such as AlphaGo or GPT-3
    - Rob Miles' Youtube channel
- Getting junior people interested in the field
    - Local EA groups, and alignment reading groups seem great here
    - Having good introductory resources, eg the Alignment Newsletter and Alignment Forum Sequences (though there's still a ton of good work to be done here!)
- Getting junior people to actually enter the field
    - MIRI does a bunch of good work here, especially their AIRCS and MSFP workshops
    - An unexpectedly important factor here is PhD supervisors willing to supervise safety students - there seems to be a major mentorship bottleneck in alignment research, with many more junior researchers excited to enter the field
        - But this seems to be increasing rapidly! Eg David Krueger is now at Cambridge and Dylan Hadfield-Menell is now at MIT

- Getting more experienced researchers to transition into the field - mostly discussed above
- Convincing key stakeholders to care about safety
    - Outreach to senior management and researchers at top AI labs, eg by having great safety teams at each
    - Talking to government and policy-makers

I emphasise all these because I think people often under-appreciate just how effective a force multiplier field-building can be, and that it eg feels much 'higher status' to be a researcher. I would argue that a large part of the value of alignment work thus far has come from field building. Though obviously a healthy field needs both!

Deconfusion Research

Deconfusion research tries to make **conceptual progress** on the key concepts related to alignment - things like intelligence, agency, goals, alignment, values, etc. Proponents tend to argue that we're hopelessly confused about what alignment even means, and have all kinds of incorrect implicit assumptions about how to view the world. And that a highly effective way to push the needle is to try to become less confused and build the relevant concepts, so we can focus on fixing the right problems and building systems that might actually work.

I most associate this view with MIRI, and found the Rocket Alignment Problem helpful for understanding this view. They draw an analogy between trying to align current systems and trying to launch a rocket to the moon without understanding Newtonian mechanics and the rocket equation. I also think some of the macrostrategy team at FHI, and Paul Christiano are doing great work here.

I found deconfusion research somewhat weird to get my head around at first - it seems ridiculously ambitious, and like it's difficult to direct well or to have real feedback loops around whether what you're doing is useful. But I've updated a lot towards this being valuable, especially when I look at successful examples of deconfusion research by alignment researchers.

I find it helpful to think about mathematics here, which contains several historical examples of *phenomenally* successful deconfusion: Probability theory is a crisp operationalisation of uncertainty. Group theory is a formal framework to reason about symmetry. Game theory is a framework to think crisply about strategy and conflict. Game Theory is a particularly interesting example, as it seems to have had a lot of work done during the cold war, by researchers deliberately trying to push the needle on ensuring cooperation between the US and USSR

Some of my favourite examples of successful deconfusion research:

- Many of the original ideas around AI Alignment - **orthogonality thesis**, **instrumental convergent goals**, **treacherous turns**, etc
    - Superintelligence is a great collection of these
- MIRI
    - Scott Garabrant and Abram Demski's work on **Embedded Agency**: In standard RL and models like AIXI we think of the agent as being

fundamentally outside of the environment and 'larger' than it. Eg AlphaGo has no notion of itself. But this cripples our ability to reason about AGI, which is fundamentally **embedded** in its environment. This will need to understand the real world, which is inherently larger than the agent, can contain copies of it, gives it the ability to self-modify, tamper with its reward function, etc.
    - Evan Hubinger's work on **mesa-optimizers** (as described above[TODO: Link to section of post])
- Work out of FHI such as Eric Drexler's **Comprehensive AI Services** (See Richard Ngo and Rohin Shah's summaries): Arguing that rather than thinking of AGI as an expected utility maximiser, we should instead imagine today's AI systems scaled up in the simplest possible way. Today's AI systems tend to be narrow tools, specific to a task. What if we automated the process of training a narrow system for a task, and solved generality that way?
- Paul Christiano (I think most posts on his blog are good examples of this, and highly worth reading!). Some of my favourites:
    - **Intent Alignment**: Arguing that the important problem to solve is creating AGI that wants to help us achieve our intentions, rather than anything more ambitious
    - **Ambitious vs Narrow Value Learning**: Distinguishing between ambitious systems, that infer our large scale goals and find creative ways to achieve them, and narrow systems, that help empower us to achieve our large scale goals by doing what we're already doing better, more like extremely competent assistants
    - **Corrigibility**: Arguing that we should aim for AGI that *wants to help us align it*. This makes the problem *much* easier, because there is a lot more room to make mistakes and fix them, rather than needing to get everything right the first time round
- Richard Ngo's AGI Safety from First Principles - a excellent and thorough case for Alignment research, revamped from the classic cases in light of what we now know about deep learning systems, and implications for what AGI might look like.

Crucial Considerations

Understanding people's threat models and research motivations, and figuring out which you do and do not agree with is a good path to get a big picture of what's going on in AI Safety. Another angle I've found useful is looking for **crucial considerations**. There are a few key, underlying beliefs about which people disagree, which often seem to generate their high-level beliefs and prioritisation. In this section, I've tried to collect some key crucial considerations, arguments for and against, and my favourite resources to read more.

Note: I tend to phrase these as binary questions. I find this a helpful framing, but in reality these are more of a spectrum, and the only reasonable thing to do is to have a distribution over where we lie on these.

Timelines

Crux: How long until we create AGI?

People seem to vary a lot on this one - some argue for <10 years (these tend to be most excited about GPT-3 and the scaling hypothesis), some for 20-50 years, some for >100 years (though I rarely hear this backed up by principled arguments)

This is extremely important for determining how much to prioritise AI Alignment as a cause area (it feels much less important if AGI is >50 years away!). It's also important for prioritising work within the field. Short timelines favour engineering work, getting really involved in our most powerful current systems, and trying to align them. Long timelines favour theoretical work, field-building and de-confusion research.

My favourite timelines work is Ajeya Cotra's report, as described above. The key to this work is the scaling hypothesis - that the main driver of progress is available compute - and carefully estimates the time until we'll have enough compute to train a model as powerful as the human brain via a range of biological anchors. The headline figure is median 30 years away (though the important thing is the framework, not the actual number).

This approach relies heavily on having a specific model of how we might get to AGI, and what the bottlenecks are. If you dislike an approach this inside-view-flavoured, you might enjoy a recent outside view focused report by Tom Davidson, which tries to model AGI via a range of reference classes, eg 'a STEM field aiming for an ambitious breakthrough', 'a transformative technology', 'a maths conjecture' etc, and estimates the probabilities from history.

Prosaic AI Alignment

Crux: Will AGI look like current systems, but bigger and better? (ie, we are not missing fundamental insights)

This is highly related to timelines, and has similar conclusions. I separate it out, because there are interestingly different arguments. If Prosaic AGI holds, then we are *much* better placed to do useful work now, and timelines are likely shorter. And you could argue that Alignment work is much more urgent, valuable and neglected in a short timelines world. So you should act *as though* Prosaic AGI was true, even if you only think there's, say, a 10% chance. And if it turns out not to be true, you can always pivot later.

See Paul Christiano's post for more details. This is definitely not obvious - for example, perhaps field building is such a big force multiplier, and needs you to start early, so you should instead assume longer timelines

Continuous vs Discontinuous Takeoff

Crux: Will our systems smoothly get more powerful between today's and AGI (continuous), or will there be a very rapid increase in capabilities (discontinuous)

This is sometimes called fast vs slow takeoff. I dislike that term, because proponents might agree that, say, we'll get AGI in 30 years, just disagree about the path there. This feels more

like a disagreement about what the world will look like 5 years before AGI than about the path we take to get there.

I found Paul Christiano's take very helpful for reasoning about this distinction, and takeoff speeds more broadly. One operationalisation might be: "Will world GDP double in 4 years before it doubles in 1 year?" (For context, world GDP currently doubles every 20-40 years - either of these would be a *big* deal)

This is very relevant for reasoning about how much capabilities will take us by surprise, and how much time we'll have to prepare. It's also key to understand the world AGI will appear in - will it be a world much like today's, or *much* weirder? A continuous takeoff proponent might be less concerned about treacherous turns because we'll probably get warning shots along the way, from systems capable enough to want to betray us, but not capable enough to pull it off well. And you could plan to use proto-AGI systems as tools to help us align AGI. While alignment feels *much* harder in a discontinuous world, where we may go from systems far dumber than us to systems far more capable without much time to adapt.

A key argument for discontinuity is the outside view: in evolutionary time, human intelligence was *very* discontinuous. You might also take inside view models, eg that a AGI could recursively self-improve, once it's smart enough to do AI research, design better algorithms, etc

A counter-argument is that evolution wasn't optimising for intelligence until fairly recently, while AI developers will be. And there are competing models of capabilities, eg that systems will be bottlenecked by hardware/compute, take a long time to train, that AI research will hit diminishing returns, etc.

How *hard* is alignment?

Crux: How hard is it going to be to align AGI? Are we mostly safe by default, with some risk of things going wrong? Or is this highly likely to go wrong without something dramatically changing?

(**Note:** The following summary is highly speculative, and I don't understand this crux that well)

This feels the vaguest crux for me, and I don't know how to operationalise this well, beyond 'what is the probability AGI causes an existential catastrophe without further longtermist intervention'. But this seems to be one of the more important drivers of disagreement.

MIRI seems one of the strongest proponents of 'alignment will be hard'. I found the post Security Mindset helpful for understanding what this perspective feels like. To understand the 'safe by default' perspective, this newsletter summarises several great interviews with researchers like Paul Christiano and Rohin Shah on a case for (relative) optimism. (Note: Both sides here still agree that Alignment should be a priority. Easy means '90% chance this goes fine', not '100% chance this goes fine')

If you believe that alignment is easy, that favours paths focused on engineering, finding ways to hack current systems into doing what we want, maybe pushing capabilities ahead on paths that seem less dangerous. It favours being more concerned about misuse risks than accident risks - caring more about who controls AGI, how power and wealth are shared in a post-AGI world, etc.

In contrast, believing that alignment is hard might lead you to be much more worried about capabilities progress at all, and making a major priority to slow things down. Anecdotally, people in this camp tend to be much more excited about theoretical, deconfusion research. The key policy focuses should be coordination to slow down deployment, and give as much time for alignment research as possible, building institutions to favour cooperation, etc. With more of a focus on irresponsible development than irresponsible deployment and accident risks over misuse risks - *any* deployment is concerning.

Conclusion

This is my current, best guess model of what the big picture of AI Alignment looks like. I'm sure many things in here are wrong, and that I'll hold quite different views 6 months from now, but I think this is much less wrong than my thoughts were 6 months ago! And I hope this was valuable for other people trying to get a sense of what the hell is going on. If you disagree with any of my summaries, or think I've missed out relevant works or perspectives, please let me know!

If you're anything like me, the natural question to ask yourself after reading this is 'what next?' There are all these different perspectives, but which one is *right*? My implicit assumption going into all this was that there would be one "correct" perspective. That most work was useless, most researchers thought everyone else's work was irrelevant, and if I thought about it hard enough, I could identify the one direction that mattered. And this felt pretty overwhelming - if all of these smart researchers can't figure out what the right way is, who do I think I'm kidding?

In practice, this seems completely wrong. Most researchers seem to think that most researchers are doing great work. They may be *more* excited about their agenda than others, but favour taking a portfolio approach to alignment work, and want everyone's work to go well. Instead, one of the biggest factors in what you work on should be your comparative advantage - find the type of work that best suits your skillset and interests, and do that (so long as the work seems broadly reasonable).

That said, I *do* think that thinking hard about research motivation and theory of change is highly worthwhile. I just don't frame it as 'finding the one true way to make a difference'. I find it easy to slip into insecurity, thinking that this problem is way too hard, and just deferring to people smarter than me about what matters, but it's *really* hard to do good research this way. To do good research, I think you need to have an inside view model of what's going on and why what you're doing matters, and to rely on that to guide what you do. This model will be wrong in many important ways, and hopefully improve over time, but it's *much* easier to do research with this model than a model built by deferring to other people, even if the second model is closer to the truth. And the way to eventually get a good model is to repeatedly try building the best model you can, accepting that it'll be wrong in many

important ways, and iteratively improving it, rather than waiting until you can build a ‘perfect’ model. (For more thoughts on this point, I recommend the section on Research in Rohin Shah’s FAQ)

If you resonate with the above, to close I recommend the following exercise: Block out an hour, and take a blank piece of paper to a quiet place without distractions. For this hour, write out as detailed a model as you can for how AI leads to a catastrophe. Pay close attention to the parts of this that feel most speculative or confusing. Then, go back through the model and poke as many holes in it as you can.