Pragmatic AI Safety

1. Introduction to Pragmatic AI Safety [Pragmatic AI Safety #1]
2. Complex Systems for AI Safety [Pragmatic AI Safety #3]
3. Perform Tractable Research While Avoiding Capabilities Externalities [Pragmatic AI Safety #4]
4. Open Problems in AI X-Risk [PAIS #5]

Introduction to Pragmatic AI Safety [Pragmatic AI Safety #1]

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

This is the introduction to a sequence of posts that describe our models for Pragmatic AI Safety. Thanks to Oliver Zhang, Mantas Mazeika, Scott Emmons, Neel Nanda, Cameron Berg, Michael Chen, Vael Gates, Joe Kwon, Jacob Steinhardt, Steven Basart, and Jacob Hilton for feedback on this sequence (note: acknowledgements here may be updated as more reviewers are added to future posts).

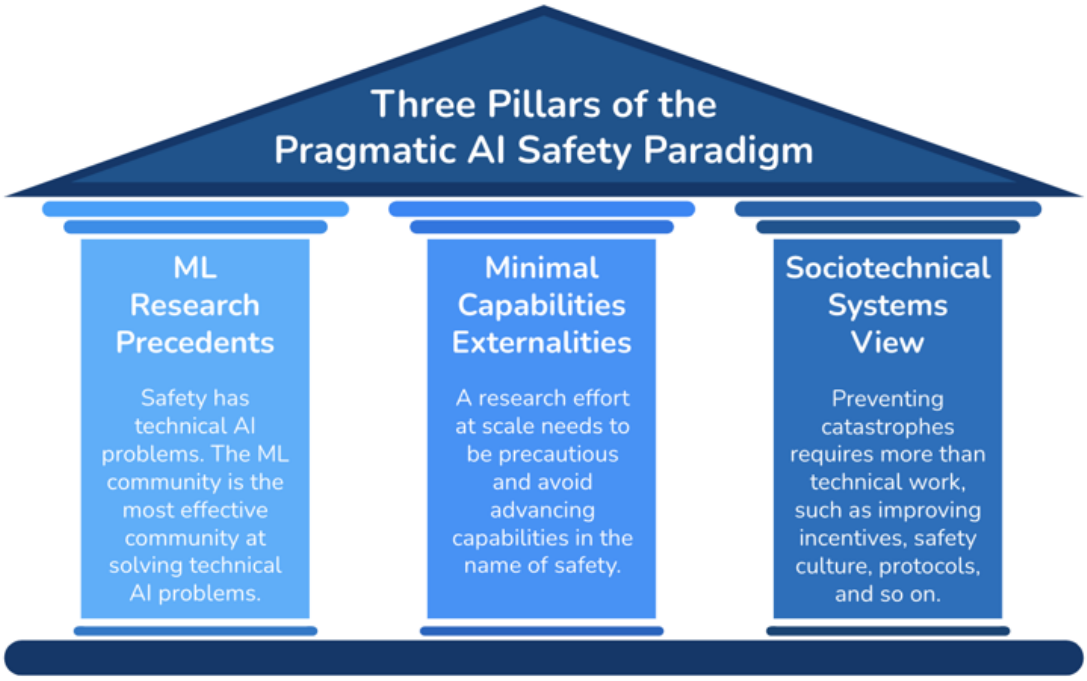
Machine learning has been outpacing safety. Ten years ago, AlexNet pushed the boundaries of machine learning, and it was trained using only two GPUs. Now state-of-the-art models are trained on thousands of GPUs. GPT-2 was released only around three years ago, and today, we have models capable of answering bar exam questions, writing code, and explaining jokes.

Meanwhile, existing approaches to AI safety have not seen similar strides. Many older approaches are still pre-paradigmatic, uncertain about what concrete research directions should be pursued and still aiming to get their bearings. Centered on math and theory, this research focuses on studying strictly futuristic risks that result from potential systems. Unfortunately, not much progress has been made, and deep learning resists the precise and universal mathematical characterizations preferred by some safety approaches.

Recently, some established safety teams have focused more on safety in the context of deep learning systems, which has the benefit of being more concrete and having faster experimental feedback loops. However, many approaches often exhibit the downside of blurring the lines between general capabilities research and safety, as there appear to be few other options.

Finally, neither the pre-paradigmatic nor industry deep learning-based approaches seriously emphasize the broad range of sociotechnical factors that are critical for reducing risk from AI systems.

Given that ML is progressing quickly, that pre-paradigmatic research is not highly scalable to many researchers, and that safety research that advances capabilities is not safely scalable to a broader research community, we suggest an approach that some of us have been developing in academia over the past several years. We propose a simple, underrated, and complementary research paradigm, which we call Pragmatic AI Safety (PAIS). By complementary, we mean that we intend for it to stand alongside current approaches, rather than replace them.



Pragmatic AI Safety rests on three essential pillars:

- *ML research precedents*. Safety involves technical AI problems, and the ML community's precedents enable it to be unusually effective at solving technical AI problems.
- *Minimal capabilities externalities. Safety research at scale needs to be precautious and avoid advancing capabilities in the name of safety.*
- *Sociotechnical systems view. Preventing catastrophes requires more than technical work, such as improving incentives, safety culture, protocols, and so on.*

ML Research Precedents

Despite relying on "broken" processes like conferences and citations, the ML community has managed to solve an increasingly general set of problems: colorizing images, protein folding, superhuman poker, art generation, etc. This doesn't mean that the ML community is set up optimally (we will discuss ways in which it's not), but it does consistently exceed our expectations and demonstrate the best track record in solving technical AI problems.

In general, ML researchers are skilled at adding arbitrary features to systems to improve capabilities, and many aspects of safety could be operationalized so as to be similarly improved. This property makes ML research precedents promising for solving technical ML problems, including many safety problems.

Here are some ML research precedents that we view as important:

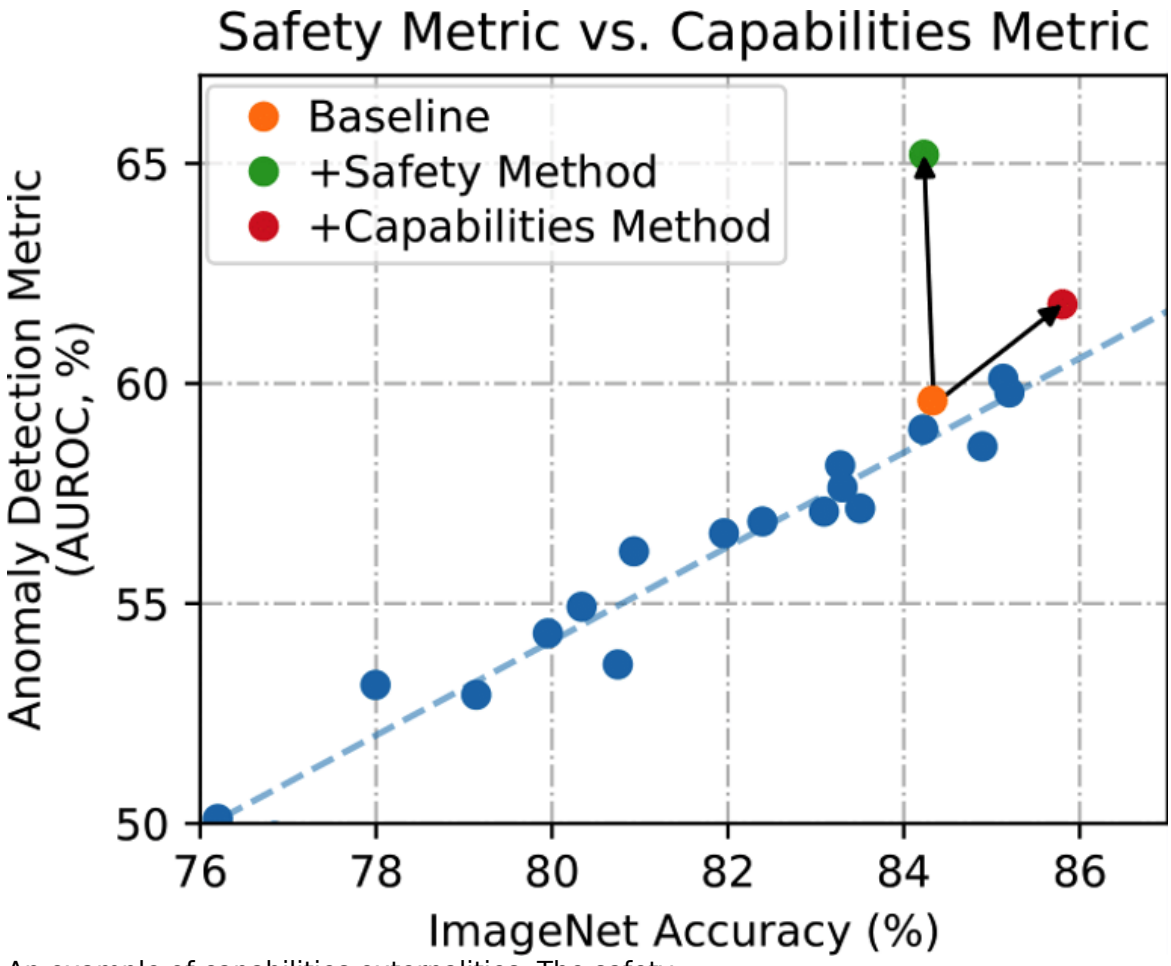
- Long term goals are broken down into empirical simplified microcosmic problems
- Subproblems can be worked on iteratively, collectively, and scalably
- Contributions are objectively measured
- The set of research priorities is a portfolio
- Researchers must convince anonymous reviewers of the value of their work
- Highly competitive, pragmatic, no-nonsense culture
- Long-run research track records are necessary for success

We will address all of these precedents throughout this sequence. They are not original to safety, and some safety researchers have been following these precedents for years in academia.

There are many problems that we consider to be included in the PAIS research umbrella that are essentially ML problems or have essential ML components: honest AI, power-averseness, implementing moral decision making, value clarification, adversarial robustness, anomaly detection, interpretable uncertainty, detection of emergent behavior, transparency, ML for cyberdefense, and ML for improved epistemics. We will cover each of these areas in depth later in the sequence.

Lastly, we should note that we consider PAIS to overlap with problems considered relevant in ML, but there are a very large number of ML problems that are not relevant to PAIS (privacy, non-convex optimization, etc.).

Minimal Capabilities Externalities



An example of capabilities externalities. The safety method makes improvements on a safety-relevant metric while avoiding progress in capabilities, while a capabilities method may improve safety simply by moving along the existing safety/capabilities curve and thus not produce much *differential* impact.

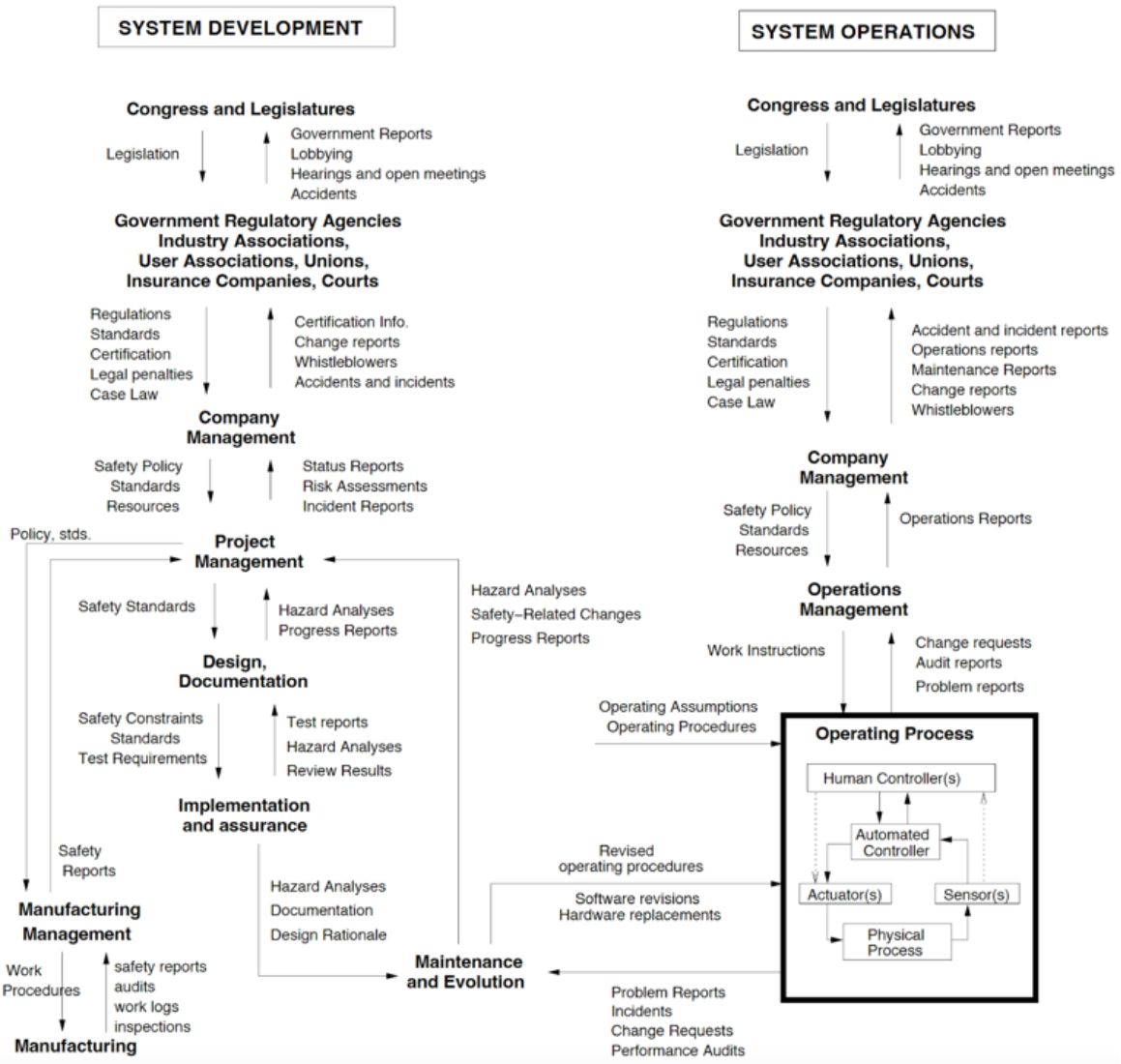
Safety and capabilities are intertwined, especially when researching deep learning systems. For example, training systems to have better world models can make them less likely to spawn unintended consequences, but also makes them more generally capable. Optimizers that can operate over longer time horizons could ensure models don't take problematic irreversible actions, but also allow models to make more complex longer-term plans in general.

It is clearly possible to make progress on safety metrics without improving capabilities, and some safety researchers have been doing it for years. But it must be done carefully. To do this, we propose a general policy of minimizing capabilities externalities. To the extent possible, we should avoid increasing capabilities in the name of safety, since this is a highly risky strategy and is not safely scalable for a broader research community. We should instead let the broader ML community take care of general capabilities, and work on safety problems that are viable with current capabilities. Rather than intuit whether something is good for overall safety, empirical researchers who follow our approach should demonstrate via measurement that their contribution does not simultaneously improve general capabilities.

Sociotechnical Systems View

A current blindspot in many AI safety approaches is to ignore nonlinear causality. Asking "how does this research agenda directly reduce this specific risk?" is well-intentioned, but it filters out accounts that capture nonlinear causal structures. Unfortunately, direct analysis is not expressive enough to model many of the most important phenomena relevant to safety. Today's interconnected systems often have nonlinear causality, including feedback loops, multiple causes, circular causation, self-reinforcing processes, butterfly effects, microscale-macroscale dynamics, and so on. There may also be emergent behavior in an overall system that cannot be attributed to any individual subcomponent.

Remote, indirect, and nonlinear causes are omitted from accounts that require linear causality. Contemporary hazard analysis, and complex systems theory in general, is aware of this deficiency, and seeks to correct it. A central takeaway from these analyses is that it is essential to consider the entire sociotechnical system when attempting to prevent failures. Rather than only focusing on the operating process (in this case, a particular AI system's technical implementation), we need to focus on systemic factors like social pressures, regulations, and perhaps most importantly, safety culture. Safety culture is one reason why engaging the broader ML community (including Chinese ML researchers) is critical, and it is currently highly underemphasized.



An example sociotechnical systems view from Leveson 2012. Much of AI safety has focused on the operating process but this process is inextricably linked to many other processes that cannot be ignored.

The Pragmatic AI Safety Sequence

In this sequence, we will describe a pragmatic approach for reducing existential risk from AI.

In the second post, which will be released alongside this post, we will present a bird's eye view of the machine learning field. Where is ML research published? What is the relative size of different subfields? How can you evaluate the credibility or predictive power of ML professors and PhD students? Why are evaluation metrics important? What is creative destruction? We will also discuss historical progress in different subfields within ML and paths and timelines towards AGI.

The third post will provide a background on complex systems and how they can be applied to both influencing the AI research field and researching deep learning.(Edit: the original third post has been split into what will now be the third and fourth posts).

The fourth post will cover problems with certain types of asymptotic reasoning and introduce the concept of capabilities externalities.

The fifth post will serve as a supplement to Unsolved Problems in ML Safety. Unlike that paper, we will explicitly discuss the existential risk motivations behind each of the areas we advocate.

The sixth and final post will focus on tips for how to conduct good research and navigate the research landscape.

A supplement to this sequence is X-Risk Analysis for AI Research.

About the authors

This sequence is being written by Thomas Woodside and Dan Hendrycks as the result of a series of conversations they've had over the last several months.

Dan Hendrycks was motivated to work exceptionally hard after reading Shelly Kagan's *The Limits of Morality* in high school. After leaving fundamentalist rural Missouri to go to college, he was advised by Bastian Stern (now at Open Philanthropy) to get into AI to reduce x-risk, and so settled on this rather than proprietary trading for earning to give. He did his undergrad in computer science; he worked on generic capabilities for a week to secure research autonomy (during which time he created the GELU activation), and then immediately shifted into safety-relevant research. He later began his PhD in computer science at UC Berkeley. For his PhD he decided to focus on deep learning rather than reinforcement learning, which most of the safety community was focused on at the time. Since then he's worked on research that defines problems and measures properties relevant for reliability and alignment. He is currently a fourth and final-year PhD student at UC Berkeley.

Thomas Woodside is a third-year undergraduate at Yale, studying computer science. Thomas did ML research and engineering at a startup and NASA before being introduced to AI safety through effective altruism. He then interned at the Center for Human-Compatible AI, working on safety at the intersection of NLP and RL. He is currently taking leave from school to work with Dan on AI safety, including working on power-seeking AI and writing this sequence.

Complex Systems for AI Safety [Pragmatic AI Safety #3]

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

This is the third post in a sequence of posts that describe our models for Pragmatic AI Safety.

It is critical to steer the AI research field in a safer direction. However, it's difficult to understand how it can be shaped, because it is very complex and there is often a high level of uncertainty about future developments. As a result, it may be daunting to even begin to think about how to shape the field. We cannot afford to make too many simplifying assumptions that hide the complexity of the field, but we also cannot afford to make too few and be unable to generate any tractable insights.

Fortunately, the field of complex systems provides a solution. The field has identified commonalities between many kinds of systems and has identified ways that they can be modeled and changed. In this post, we will explain some of the foundational ideas behind complex systems and how they can be applied to shaping the AI research ecosystem.

Along the way, we will also demonstrate that deep learning systems exhibit many of the fundamental properties of complex systems, and we show how complex systems are also useful for deep learning AI safety research.

A systems view of AI safety

Background: Complex Systems

When considering methods to alter the trajectory of empirical fields such as deep learning, as well as preventing catastrophe from higher risk systems, it is necessary to have some understanding of complex systems. Complex systems is an entire field of study, so we cannot possibly describe every relevant detail here. In this section, we will try merely to give a very high level overview of the field. At the end of this post we present some resources for learning more.

Complex systems are systems consisting of many interacting components that exhibit emergent collective behavior. Complex systems are highly interconnected, making decomposition and reductive analysis less effective: breaking the system down into parts and analyzing the parts cannot give a good explanation of the whole. However, complex systems are also too organized for statistics, since the interdependencies in the system break fundamental independence assumptions in much of statistics. Complex systems are ubiquitous: financial systems, power grids, social insects, the internet, weather systems, biological cells, human societies, deep learning models, the brain, and other systems are all complex systems.

It can be tricky to compare AI safety to making other specific systems safer. Is making AI safe like making a rocket, power plant, or computer program safe? While analogies can be found, there are many disanalogies. It's more generally useful to talk about making complex systems safer. For systems theoretic hazard analysis, we can abstract away from the specific content and just focus on shared structure across systems. Rather than talk about what worked well for one high-risk technology, with a systems view we can talk about what worked well for a large number of them, which prevents us from overfitting to a particular example.

The central lesson to take away from complex systems theory is that reductionism is not enough. It's often tempting to break down a system into isolated events or components, and then try to analyze each part and then combine the results. This incorrectly assumes that separation does not distort the system's properties. In reality, parts do not operate independently, and are subject to feedback loops and nonlinear interactions. Analyzing the pairwise interactions between parts is not sufficient for capturing the full system complexity (this is partially why a bag of n-grams is far worse than attention).

Hazard analysis once proceeded by reductionism alone. In earlier models, accidents are broken down into a chain of events thought to have caused that accident, where a hazard is a root cause of an accident. Complex systems theory has supplanted this sort of analysis across many industries, in part because the idea of an ultimate "root cause" of a catastrophe is not productive when analyzing a complex system. Instead of looking for a single component responsible for safety, it makes sense to identify the numerous factors, including sociotechnical factors, that are contributory. Rather than break events down into cause and effect, a systems view instead sees events as a product of a complex interaction between parts.

Recognizing that we are dealing with complex systems, we will now discuss how to use insights from complex systems to help make AI systems safer.

Improving Contributing Factors

"Direct impact," that is impact produced from a simple, short, and deterministic causal chain, is relatively easy to analyze and quantify. However, this does not mean that direct impact is always the best route to impact. If someone only focuses on direct impact, they won't optimize for diffuse paths towards impact. For instance, EA community building is indirect, but without it there would be far fewer funds, fewer people working on certain problems, and so on. Becoming a billionaire and donating money is indirect, but without this there would be significantly less funding. Similarly, safety field-building may not have an immediate direct impact on technical problems, but it can still vastly change the resources devoted to solving those problems, in turn contributing to solving them (note that "resources" does not (just) mean money, but rather competent researchers capable of making progress). In a complex system, such indirect/diffuse factors have to be accounted for and prioritized.

AI safety is not all about finding safety mechanisms, such as mechanisms that could be added to make superintelligence completely safe. This is a bit like saying computer security is all about firewalls, which is not true. Information assurance evolved to address blindspots in information security, because it is understood that we cannot ignore complex systems, safety culture, protocols, and so on.

Often, research directions in AI safety are thought to need to have a simple direct impact story: if this intervention is successful, what is the short chain of events towards it being useful for safe and aligned AGI? "How does this directly reduce x-risk" is a well-intentioned question, but it leaves out salient remote, indirect, or nonlinear causal factors. Such diffuse factors cannot be ignored, as we will discuss below.

A note on tradeoffs with simple theories of impact

AI safety research is complex enough that we should expect that understanding a theory of impact might require deep knowledge and expertise about a particular area. As such, a theory of impact for that research might not be easily explicable to somebody without any background in a short amount of time. This is especially true of theories of impact that are multifaceted, involve social dynamics, and require an understanding of multiple different angles of the problem. As such, we should not only focus on theories of impact that are easily explicable to newcomers.

In some cases, pragmatically one should not always focus on the research area that is most directly and obviously relevant. At first blush, reinforcement learning (RL) is highly relevant to advanced AI agents. RL is conceptually broader than supervised learning such that supervised learning can be formulated as an RL problem. However, the problems considered in RL that aren't considered in supervised learning are currently far less tractable. This can mean that in practice, supervised learning may provide more tractable research directions.

However, with theories of impact that are less immediately and palpably related to x-risk reduction, we need to be very careful to ensure that research remains relevant. Less direct connection to the essential goals of the research may cause it to drift off course and fail to achieve its original aims. This is especially true when research agendas are carried out by people who are less motivated by the original goal of the research, and could potentially lead to value drift where previously x-risk-motivated researchers become motivated by proxy goals that are no longer relevant. This means that it is much more important for x-risk-motivated researchers and grantmakers to maintain the field and actively ensure research remains relevant (this will be discussed later).

Thus, there is a tradeoff involved in only selecting immediately graspable impact strategies. Systemic factors cannot be ignored, but this does not eliminate the need for understanding causal (whether indirect/nonlinear/diffuse or direct) links between research and impact.

Examples of the importance of systemic factors

The following examples illustrate the extreme importance of systemic factors (and the limitations of direct causal analysis and complementary techniques such as backchaining):

- Increasing wealth is strongly associated with a reduction in childhood mortality. But one cannot always credit the survival of any particular child to an increase of the wealth of their country. Nonetheless, a good way to reduce childhood mortality is still to increase overall wealth.
- Community building, improving institutions, and improving epistemics can usually not be linked directly to specific outcomes, but in aggregate they clearly have large effects.
- Smoking does not guarantee you will get cancer. If you smoke and get cancer, it is not necessarily because you smoked. Still, avoiding smoking is clearly a good way to avoid cancer. Contrariwise, exercise does not guarantee that you will be healthy, but it robustly helps.
- Intelligence (e.g. as measured by IQ) has an enormous impact on the ability of people to perform various tasks. But it is implausible to point to a particular multiple choice test question that somebody answered correctly and say "they got this question because their IQ was above x." Similarly, forecasting and rationality could increase the "IQ" of the superorganism, but it similarly could not be expected to produce one single definite outcome. Improving the rationality waterline helps with outcomes, even if we cannot create a simple chain of events showing that it will prevent a particular future catastrophe.
- Any particular hurricane or wildfire cannot be attributed to the effects of climate change, but reducing climate change is a good way to reduce the prevalence of those extreme weather events.

In the cases above, it is possible to use statistics to establish the relationship between the variables given enough data. Some can be causally established through randomized controlled trials. However, we do not have the ability or time to run an RCT on diffuse factors that reduce x-risk from AI. Unlike the situations above, we do not get to observe many different outcomes because an existential catastrophe would be the last observation we would make. This does not mean diffuse factors are unimportant; on the contrary, they are extremely important. We can instead identify time-tested factors that have been robustly useful in similar contexts in the past.

On a more societal scale, the following diffuse factors are quite important for reducing AI x-risk. Note that these factors may interact in some cases: for instance, proactivity about risks might not help much if malevolent actors are in power.

- **People having improved epistemics:** Irrationality could cause people to ignore warning signs, dismiss correct claims, and barrel ahead when they shouldn't.
- **Proactivity about (tail) risks:** Causing humanity as a collective to care more about tail risks would be a boon for safety. Work on mitigating tail risks is currently underincentivized due to the human tendency to ignore tail risks.
- **Expanded moral circles:** The term "moral circle" describes the beings that one considers to be morally relevant (e.g. people in your community, people across the world, future people, non-human animals, etc.). People do not need a large moral circle to want to avoid their own deaths, but it can strengthen the perceived importance of reducing x-risk.
- **Keeping (misaligned) malevolent actors (egoists/Machiavellians/psychopaths) out of power:** Contending with actively malevolent leaders is even more difficult than contending with apathetic leaders. Getting even-handed, cautious, and altruistic people into positions of power is likely to reduce x-risk.

Sociotechnical Factors

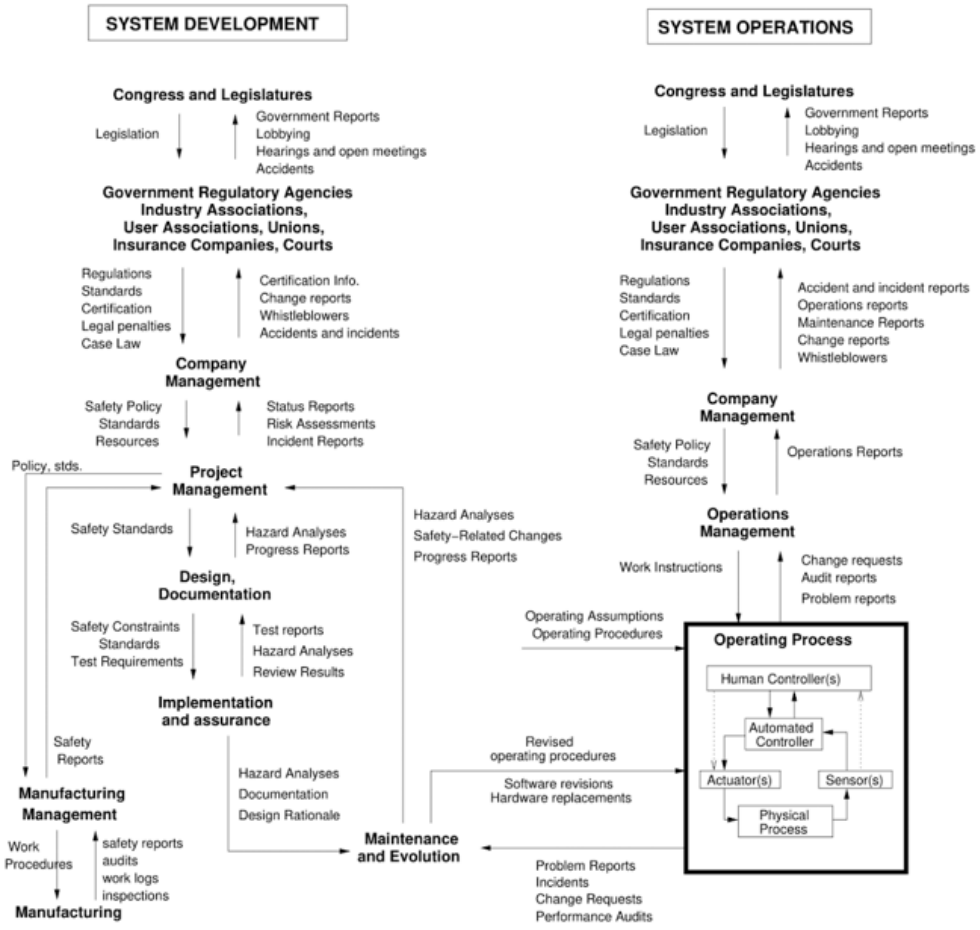


Figure 4.4: General Form of a Model of Socio-Technical Control

An abstract template from Nancy Leveson illustrating the complex interplay between sociotechnical factors and an operating process.

We can now speak about specific diffuse factors that have shown to be highly relevant to making high-risk technological systems safer, which are also relevant to making present and future AI systems safer. The following sociotechnical factors (compiled from Perrow, La Porte, Leveson, and others) tend to influence hazards:

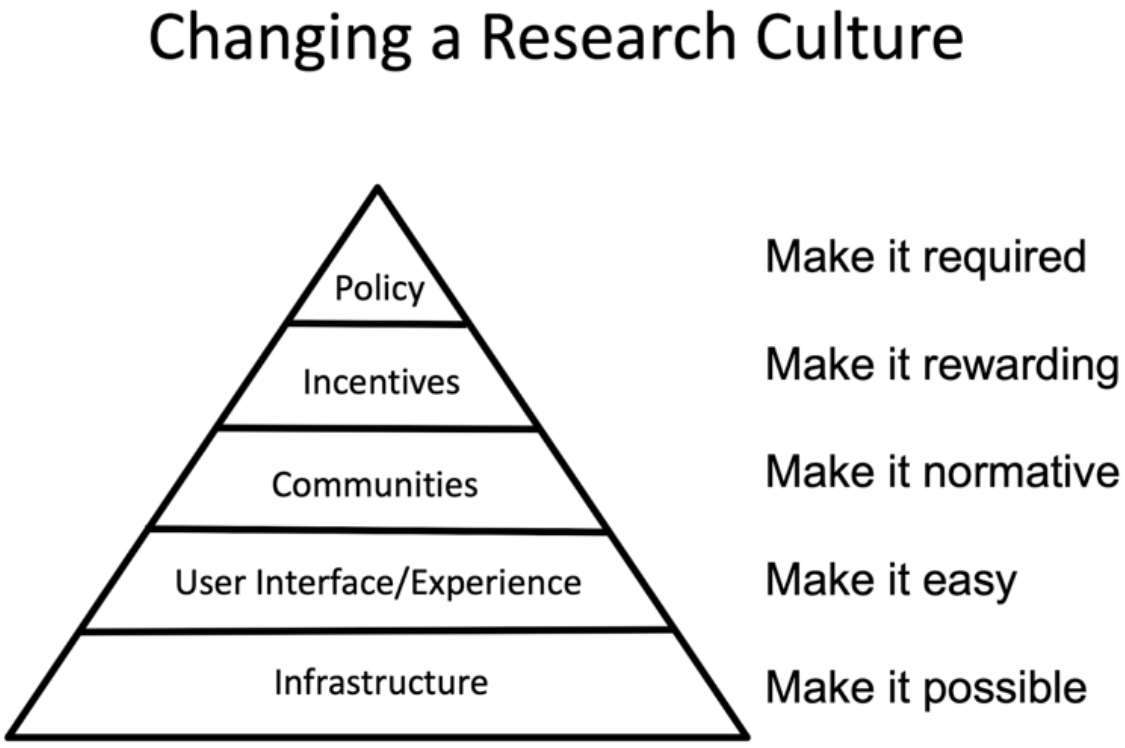
- **Rules and regulations**, perhaps including internal policies as well as legal governance.
- **Social pressures**, including those from the general public as well as well-connected powerful people.
- **Productivity pressures**, or pressure to deliver quickly.
- **Incentive structures** within the organization, such as benefits to delivering quickly or retaliation for whistleblowing.
- **Competition pressures from other actors** who may have different safety standards, or otherwise be able to move faster.
- **Safety budget and compute allocation**: are safety teams capable of running the experiments they need to? Is a significant proportion of the budget and compute dedicated to safety?
- **Safety team size**, which is related to budget. The number of researchers, engineers, and top researchers on the safety team matters a lot.
- **Alarm fatigue**: if many false alarms are raised about safety issues which were never borne out, this could reduce willingness to care about safety.
- **Reduction in inspection and preventative maintenance**, which is perhaps less relevant for a forward-looking problem like safety. However, if people do not keep a close eye on capabilities, this could allow for emergent capabilities (or actors) to take us by surprise.
- **Lack of defense in depth**: overlapping systems that provide multiple layers of defense against hazards.
- **Lack of redundancy**: multiple systems which accomplish similar safety tasks, so as to remove single points of failure.
- **Lack of fail-safes**: features that allow a system to fail gracefully.
- **Safety mechanism cost**: how much does it cost to make a system safe?
- **Safety culture**, or the general attitude towards safety within an organization or field.

According to Leveson, who has been consulted on the design of high-risk technologies across numerous industries, "*the most important [contributing factor] to fix if we want to prevent future accidents*" is safety culture.

Safety Culture

Safety culture is not an easy risk factor to address, though it is likely to be one of the most important. Many ML researchers currently roll their eyes when asked about alignment or safety: usually, one cannot simply go straight to discussing existential risks from superintelligences without suffering social costs or efforts potentially backfiring. This is a sign of a deficient safety culture.

How do we improve safety culture? Safety needs to be brought to the forefront through good incentive structures and serious research. Pushing research cultures in a safer direction is bottlenecked by finding interesting, shovel-ready, safety-relevant tasks for people to do and funding them to complete those tasks.



As suggested by the speculative pyramid above, it is not realistic to immediately try to make safety into a community norm. Before this can be done, we need to make it clear what safety looks like and we need infrastructure to make AI safety research as easy as possible. Researchers need to accept arguments about risks *and* they need clear, concrete, low-risk research tasks to pursue. This involves creating funding opportunities, workshops, and prizes, as well as clearly defining problems through metrics.

Some contributing factors that can improve safety culture are as follows:

- **Preoccupation with failure**, especially black swan events and unseen failures.
- **Reluctance to simplify interpretations** and explain failures using only simplistic narratives.
- **Sensitivity to operations**, which involves closely monitoring systems for unexpected behavior.
- **Commitment to resilience**, which means being rapidly adaptable to change and willing to try new ideas when faced with unexpected circumstances.
- **Under-specification of organizational structures**, where new information can travel throughout the entire organization rather than relying only on fixed reporting chains.

For mainstream culture, public outreach can help. One plausible way that AI systems could become more safe is due to a broader cultural desire for safety, or fear of lack of safety. Conversely, if AI safety is maligned or not valued in the general public, there may be other public pressures (e.g. winning the AI race, using AI to achieve some social good quickly) that could push against safety. Again, mainstream outreach should not be so extreme as to turn the research community against safety. Overton windows must be shifted with care.

Currently, safety is being attacked by critics who believe that it detracts from work on AI fairness and bias and does not heavily prioritize current power inequalities, which they view

as the root cause of world problems. Criticisms have been connected to criticisms of longtermism, particularly absurd-seeming expected value calculations of the number of future beings, as well as the influence of EA billionaires. These criticisms threaten to derail safety culture. It is tricky but necessary to present an alternative perspective while avoiding negative side effects.

Some technical problems are instrumentally useful for safety culture in addition to being directly useful for safety. One example of this is reliability: building highly reliable systems trains people to specifically consider the tail-risks of their system, in a way that simply building systems that are more accurate in typical settings does not. On the other hand, value learning, while it is also a problem that needs to be solved, is currently not quite as useful for safety culture optimization.

Composition of top AI researchers

We will now discuss another contributing factor that is important to improve: the composition of top AI researchers. In the future, experimenting with the most advanced AI systems will be extraordinarily expensive (in many cases, it already is). A very small number of people will have the power to set research directions for these systems. Though it's not possible to know exactly who will be in this small group, it could comprise any number of the top AI researchers today. However, one thing is known: most top AI researchers are not sympathetic to safety. Consequently, there is a need to increase the proportion of buy-in among top researchers, especially including researchers in China, and also to train more safety-conscious people to be top researchers.

It's tempting to think that top AI researchers can simply be bought. This is not the case. To become top researchers, they had to be highly opinionated and driven by factors other than money. Many of them entered academia, which is not a career path typically taken by people who mainly care about money. Yann LeCun and Geoffrey Hinton both still hold academic positions in addition to their industry positions at Meta and Google, respectively. Yoshua Bengio is still in academia entirely. The tech companies surely would be willing to buy more of their time for a higher price than academia, so why are the three pioneers of deep learning not all in the highest-paying industry job? Pecuniary incentives are useful for externally motivated people, but many top researchers are mostly internally motivated.

As discussed in the last post, a leading motivation for researchers is the interestingness or "coolness" of a problem. Getting more people to research relevant problems is highly dependent on finding interesting and well-defined subproblems for them to work on. This relies on concretizing problems and providing funding for solving them.

Due to the fact that many top researchers are technopositive, they are not motivated by complaints about the dangers of their research, and they are likely to be dismissive. This is especially true when complaints come from those who have not made much of a contribution to the field. As a result, it is important to keep the *contribution to complaint ratio* high for those who want to have any credibility. "Contribution" can be a safety contribution, but it needs to be a legible contribution to ML researchers. Top researchers may also associate discussion of existential risk with sensationalist stories in the media, doom-and-gloom prophecies, or panic that "we're all going to die."

Causes of Neglectedness

There are a number of additional factors which contribute to the general neglectedness of AI safety. It is important to optimize many of these factors in order to improve safety. A more general list of these factors is as follows.

- **Corporate**: myopic desire for short-term shareholder returns, safety features may take a long time to pay off, some human values may be difficult to incorporate in prices or pecuniary incentives

- **Temperamental**: techno-optimism, distaste for discussing risks
- **Political**: AI safety is seen to compete with more politically popular causes like climate change and reducing inequality
- **Technical Background**: safety problems are outside of one's existing skill set or training, and likewise machine ethics and sociotechnical concerns and do not as easily as easily comport with their quantitative inclinations
- **Socioeconomic distance**: many AI researchers live in tech bubbles, which can cause researchers to devalue or implicitly underemphasize cosmopolitan approaches towards loading human values
- **Tail risks:** highly consequential black swans and tail risks are systematically neglected
- **Respectability**: distaste for talk of AGI, feeling an area is not prestigious, areas associated with people who hold other unpopular or weird-seeming ideas
- **Temporal**: future risks and future people are highly neglected

Complex Systems for AI Safety

Complex systems studies emphasizes that we should focus on contributing factors (as events are the product of the interaction of many contributing factors), and it helps us identify which contributing factors are most important across many real-world contexts. They also provide object-level insight about deep learning, since deep learning systems are themselves complex systems.

Deep learning exhibits many halmarks of complex systems:

- *Highly distributed functions*: partial concepts are encoded redundantly and highly aggregated
- *Numerous weak nonlinear connections*: Connection parameters are nonzero (rather than sparse) and neural networks contain nonlinear activation functions
- *Self-organization*: optimizing a loss automatically specifies a model's internal content
- *Adaptivity*: few-shot models and online models are adaptive
- *Feedback loops*: Self-play, human in the loop, auto-induced distribution shift
- *Scalable structure*: scaling laws show that models scale simply and consistently
- *Emergent capabilities*: numerous unplanned capabilities spontaneously "turn on"

As such, insights from complex systems are quite applicable to deep learning. Similarly, like all large sociotechnical structures, the AI research community can also be considered to be a complex system. The organizations operating AI systems are also complex systems.

Complex systems is a *predictive*—not just explanatory—model for various problems, including AI safety. In fact, many important concepts in AI safety turn out to be specific instances of more general principles. Here are examples of *highly simplified* lessons from complex systems, mostly from The Systems Bible (1975):

- **Systems develop goals of their own the instant they come into being.**
  - *Explanation:* A system's goal is seldom merely the initial goal it was tasked with. Rather, other goals emerge from the organization of the system.
  - *Implications for AI:* One salient example are instrumental goals for self-preservation or power-seeking.
- **Intrasystem goals come first.**
  - *Explanation:* Systems often decompose goals into subparts for different intrasystem components to solve. During this decomposition, goals are often distorted. A common failure mode is that the system's explicitly written objective is not necessarily the objective that the system operationally pursues, and this can result in misalignment. A system's subgoals can supersede its actual goals. For example, a bureaucratic department (a subsystem) can capture power and have the company pursue goals unlike its original goals.
  - *Implications for AI:* A related phenomenon is already well known to the community as mesa-optimization; it has been predicted on a more general level

by systems theory for decades.

- **The mode of failure of a complex system cannot ordinarily be predicted from its structure.**
  - *Explanation:* Simply examining a complex system will not necessarily give you a good idea for how it might fail. Failures are usually identified from experience and testing.
  - *Implications for AI:* It is difficult to understand how all the ways a neural network might fail simply by examining its weights or architecture or through armchair/whiteboard analysis. We can count on some failures being unpredictable. (Although failures are inevitable, catastrophes are not.)
  - *Implications for strategy:* An approach of "think about the problem really hard and make sure there are no holes in the solution" is unlikely to turn up a solution that truly has no holes. Preventing failure in a complex system is not a math problem. In complex systems there are few symmetries, few necessary and sufficient conditions or boolean connectives (no root cause), circular relationships, numerous partial concepts (combinatorial explosion), self-organization, high distributivity. All of these properties make complex systems very difficult to analyze from an armchair/whiteboard or with proofs.
- **The crucial variables are discovered by accident.**
  - *Explanation*: It is difficult to know what the most important parts of a system are by inspection. The highest points of leverage are not obvious. Likewise, the methods that will work best are often found by tinkering or serendipity.
  - *Implications for AI:* Many of the greatest breakthroughs in AI are not discovered purely by principled, highly structured investigation, but instead by tinkering.
  - *Implications for strategy*: Many current approaches to research bet on AGI being best represented as a mathematical object rather than a complex system, which seems unrealistic given current AI systems as well as other intelligent systems we know (e.g. humans, corporations).
- **A large system, produced by expanding the dimensions of a smaller system, does not behave like the smaller system.**
  - *Explanation:* Purely scaling up a system does not only make it better at whatever it was doing before. We should expect to see new qualitative properties and emergent capabilities.
  - *Implications for AI:* We should expect to also see emergent capabilities that did not exist at all in smaller versions. For example, at low levels of capabilities, deception is not a good idea for an intelligence, but as it becomes more intelligent, deception may be a better strategy for achieving goals.
  - *Implications for strategy:* Scaling up an aligned system and expecting it to be fully aligned is not an airtight idea. Scaling, even of a highly reliable system, needs to be done carefully.
- **(From Gilb) Gilb's Laws of Unreliability: any system which depends on human reliability is unreliable.**
  - *Explanation:* Humans are not reliable. Reliance on them will create unreliability.
  - *Implications for strategy:* AI systems may be too explosive and fast-moving for depending heavily on human feedback or human-in-the-loop methods. We will need a more reliable strategy for preserving human values, perhaps through oversight from other AI systems.
- **A complex system that works is invariably found to have evolved from a simple system that works.**
  - *Explanation:* Complex systems cannot be created from scratch and expected to work. Rather, they have to evolve from simpler functioning systems.
  - *Implications for strategy:* Working on safety for simpler systems, and attempting to (carefully) scale them up is more likely to be successful than starting by trying to build an aligned complex system from scratch. Although systems behave differently when scaled, the ones that work are evolved from smaller systems. If one is unable to align a simpler version of a complex system, it is unlikely that one can align the more complex version. On this view a top priority is making today's simpler systems safer.

Diversification

There are many different facets involved in making complex systems work well; we cannot simply rely on a single contributing factor or research direction. The implication is that it makes sense to diversify our priorities.

Since an individual has limited ability to become specialized and there are many individuals, it often makes sense to bet on the single highest expected value (EV) research approach. However, it would be a mistake to think of the larger system in the same way one thinks of an individual within the system. If the system allocates all resources into the highest EV option, and that sole option does not pay off, then the system fails. This is a known fact in finance and many other fields that take a portfolio approach to investments. Do not make one big bet or only bet on the favorite (e.g., highest estimated EV) avenue. The factor with the highest return on investment in isolation is quite different from the highest return on investment *profile* spanning multiple factors. The marginal benefit of X might be higher than Y, but the system as a whole is not forced to choose only one. As the common adage goes, "don't put all your eggs in one basket."

One example of obviously suboptimal resource allocation is that the AI safety community spent a very large fraction of its resources on reinforcement learning until relatively recently. While reinforcement learning might have seemed like the most promising area for progress towards AGI to a few of the initial safety researchers, this strategy meant that not many were working on deep learning. Deep learning safety researchers were encouraged to focus on RL environments because it is "strictly more general," but just because one can cast a problem as a reinforcement learning problem does not mean one should. At the same time, the larger machine learning community focused more on deep learning than reinforcement learning. Obviously, deep learning appears now to be at least as promising as reinforcement learning, and a lot more safety research is being done in deep learning. Due to tractability, the value of information, iterative progress in research, and community building effects, it might have been far better had more people been working on deep learning from an earlier date. This could readily have been avoided had the community leaders heeded the importance of heavily diversifying research.

If we should address multiple fronts simultaneously, not bet the community on a single area or strategy, we will pay lower costs from neglecting important variables. Since costs often scale superlinearly with the time a problem has been neglected, which has serious practical implications, it makes sense to apply resources to pay costs frequently, rather than only applying resources after costs have already blown up. The longer one waits, the more difficult it could be to apply an intervention, and if costs are convex (e.g. quadratic rather than logarithmic), costs are exacerbated further. Diversification implicitly keeps these costs lower.

AI safety is an area with extremely high uncertainty: about what the biggest problems will be, what timelines are, what the first AGI system will look like, etc. At the highest levels of uncertainty, it is most important to *improve the virtues of the system* (e.g., meritocratic structures, sheer amount of talent, etc.). If your uncertainty level is slightly less, you *additionally* want to make a few big bets and numerous small bets created in view of a range of possible futures. Moreover, under high uncertainty or when work is inchoate, it is far more effective to follow an "emergent strategy," not define the strategy with a highly structured, perfected direction.

With diversification, we do not need to decisively resolve all of the big questions before acting. Will there be a slow takeoff, or will AI go foom? Are the implicit biases in SGD beneficial to us, or will they work against us? Should we create AI to pursue a positive direction, or should we just try to maximize control to prevent it from taking over? So long as answers to these questions are not highly negatively correlated, we can diversify our bets and support several lines of research. Additionally, research can help resolve these questions

and can inform which future research should be included in the overall portfolio. Seeing value in diversification saves researchers from spending their time articulating their tacit knowledge and highly technical intuitions to win the court of public opinion, as perhaps the question cannot be resolved until later. Diversification makes researchers less at odds with each other and lets them get on with their work, and it reduces our exposure to risks from incorrect assumptions.

Diversification does not mean that one should not be discretionary about ideas. Some ideas, including those commonly pursued in academia and industry, may not be at all useful to x-risk, even if they are portrayed that way. Just because variables interact nonlinearly does not mean that resources should be devoted to a variable that is not connected with the problem.

In addition, *individuals* do not necessarily need to have a diverse portfolio. There is a benefit to specialization, and so individuals may be better off choosing a single area where they are likely to reach the tail of impact through specialization. However, if everyone individually focused on what they viewed as the most important area of research overall, and their judgments on this were highly correlated, we would see a concentration of research into only a few areas. This would lead to problems, because even if these areas are the most important, they should not be single-mindedly pursued to the neglect of all other interventions.

In complex systems, we should expect many multiplicatively interacting variables to be relevant to the overall safety of a system (we will discuss this model more in the next post). If we neglect other safety factors only to focus on "the most important one," we are essentially setting everything else to zero, which is not how one reduces the probability of risk in a multiplicative system. For instance, we should not just focus on creating technical safety solutions, let alone betting on one main technical solution. There are other variables that can be expected to nonlinearly interact with this variable: the cost of such a system, the likelihood of AGI being developed in a lab with a strong safety culture, the likelihood of other actors implementing an unaligned version, and the likelihood of the aligned system in question being the one that actually leads to AGI. These interactions and interdepencies imply that effort must be expended to push on all factors simultaneously. This can also help provide what is called *defense in depth*: if one measure for driving down x-risk fails, other already existing measures can help handle the problem.

Like many outcomes, impact is long tailed, and the impact of a grant will be dominated by a few key paths to impact. Likewise, in a diverse portfolio, the vast majority of the impact will likely be dominated by a few grants. However, the best strategies will *sample heavily from the long tail distribution*, or maximize exposure to long tail distributions. Some ways to increase exposure to the black swans are with broad interventions that could have many different positive impacts, as well as a larger portfolio of interventions. This contrasts with an approach that attempts to select only targeted interventions in the tails, which is often infeasible in large, complex systems because the tails cannot be fully known beforehand. Instead, one should prioritize interventions that have a sufficient chance of being in the tails.

Depending on what phase in the development of AI we are, targeted or broad interventions should be more emphasized in the portfolio. In the past, broad interventions would clearly have been more effective: for instance, there would have been little use in studying empirical alignment prior to deep learning. Even more recently than the advent of deep learning, many approaches to empirical alignment were highly deemphasized when large, pretrained language models arrived on the scene (refer to our discussion of creative destruction in the last post). Since the deep learning community is fairly small, it is relatively tractable to work on broad interventions (in comparison to e.g. global health, where interventions will need to affect millions of people).

At this stage, targeted interventions to align particular systems are not currently likely to deliver all the impact, nor are broad approaches that hope to align *all* possible systems. This is because there is still immense upside in optimizing contributing factors to good research,

which will in turn cause both of these approaches to be dramatically more effective. The best interventions will look less like concrete stories for how the intervention impacts a particular actor during the creation of AGI and more like actions that help to improve the culture/incentives/buy-in of several possible actors

This suggests that a useful exercise might be coming up with broad interventions that equip the safety research field to deal with problems more effectively and be better placed to deliver targeted interventions in the future. Note that some broad interventions, like interventions that affect safety culture, are not simply useful insofar as they accelerate later targeted interventions, but also in that they may increase the likelihood of those targeted interventions being successfully adopted.

We also need to have targeted interventions, and they may need to be developed before they are known to be needed due to the risk of spontaneously emergent capabilities. There is also an argument that developing targeted interventions now could make it easier to develop targeted interventions in the future. As a result, a mixture of targeted and broad interventions is needed.

Conclusion

It can be daunting to even begin to think how to influence the AI research landscape due to its size and complexity. However, the study of complex systems illuminates some common patterns that can help make this question more tractable. In particular, in many cases it makes more sense to focus on improving contributing factors rather than only try to develop a solution that has a simple, direct causal effect on the intended outcome. Complex systems are also useful for understanding machine learning safety in general, since both the broader research community, deep learning systems, and the organizations deploying deep learning systems are all complex systems.

Resources on Complex Systems

Complex systems is a whole field of study that can't possibly be fully described in this post. We've added this section with resources for learning more.

- (If you only look at one, look at this:) An introduction to contemporary hazard analysis that justifies the methods far more completely than this post can.
- A short video introduction to complex systems.
- A short video introduction to emergence, a key property of complex systems.
- Systemantics by John Gall, one of the foundational texts of complex systems.
- A class introduction to complex systems.

Perform Tractable Research While Avoiding Capabilities Externalities [Pragmatic AI Safety #4]

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

This is the fourth post in a sequence of posts that describe our models for Pragmatic AI Safety.

We argued in our last post that the overall AI safety community ought to pursue multiple well-reasoned research directions at once. In this post, we will describe two essential properties of the kinds of research that we believe are most important.

First, we want research to be able to tractably produce tail impact. We will discuss how tail impact is created in general, as well as the fact that certain kinds of asymptotic reasoning exclude valuable lines of research and bias towards many forms of less tractable research.

Second, we want research to avoid creating capabilities externalities: the danger that some safety approaches produce by way of the fact that they may speed up AGI timelines. It may at first appear that capabilities are the price we must pay for more tractable research, but we argue here and in the next post that these are easily avoidable in over a dozen lines of research.

Strategies for Tail Impact

It's not immediately obvious how to have an impact. In the second post in this sequence, we argued that research ability and impact is tail distributed, so most of the value will come from the small amount of research in the tails. In addition, trends such as scaling laws may make it appear that there isn't a way to "make a dent" in AI's development. It is natural to fear that the research collective will wash out individual impact. In this section, we will discuss high-level strategies for producing large or decisive changes and describe how they can be applied to AI safety.

Processes that generate long tails and step changes

Any researcher attempting to make serious progress will try to maximize their probability of being in the tail of research ability. It's therefore useful to understand some general mechanisms that tend to lead to tail impacts. The mechanisms below are not the only ones: others include thresholds (e.g. tipping points and critical mass). We will describe three processes for generating tail impacts: multiplicative processes, preferential attachment, and the edge of chaos.

Multiplicative processes

Sometimes forces are additive, where additional resources, effort, or expenditure in any one variable can be expected to drive the overall system forward in a linear way. In cases like this, the Central Limit Theorem often holds, and we should expect that

outcomes will be normally distributed–in these cases one variable tends not to dominate. However, sometimes variables are multiplicative or interact nonlinearly: if one variable is close to zero, increasing other factors will not make much of a difference.

In multiplicative scenarios, outcomes will be dominated by the combinations of variables where each of the variables is relatively high. For example, adding three normally distributed variables together will produce another normal distribution with a higher variance; multiplying them together will produce a long-tailed distribution.

As a concrete example, consider the impact of an individual researcher with respect to the variables that impact their work: time, drive, GPUs, collaborators, collaborator efficiency, taste/instincts/tendencies, cognitive ability, and creativity/the number of plausible concrete ideas to explore. In many cases, these variables can interact nonlinearly. For example, it doesn't matter if a researcher has fantastic research taste and cognitive ability if they have no time to pursue their ideas. This kind of process will produce long tails, since it is hard for people to get all of the many different factors right (this is also the case in startups).

The implication of thinking about multiplicative factors is that we shouldn't select people or ideas based on a single factor, and should consider a range of factors that may multiply to create impact. For instance, selecting researchers purely based on their intelligence, mathematical ability, programming skills, ability to argue, and so on is unlikely to be a winning strategy. Factors such as taste, drive, and creativity must be selected for, but they take a long time to estimate and are often revealed through their long-term research track record. Some of these factors are less learnable than others, so consequently it may not be possible to become good at all of these factors through sheer intellect or effort given limited time.

Multiplicative factors are also relevant in the selection of *groups* of people. For instance, in machine learning, selecting a team of IMO gold medalists may not be as valuable as a team that includes people with other backgrounds and skill sets. The skillsets of some backgrounds have skill sets which may cover gaps in skill sets of people from other backgrounds.

Preferential Attachment

In our second post, we addressed the Matthew Effect: *to those who have, more will be given.* This is related to a more general phenomenon called preferential attachment. There are many examples of this phenomenon: the rich get richer, industries experience agglomeration economies, and network effects make it hard to opt out of certain internet services. See a short video demonstrating this process here. The implication of preferential attachment and the Matthew Effect is that researchers need to be acutely aware that it helps a lot to do very well early in their careers if they want to succeed later. Long tail outcomes can be heavily influenced by timing.

Edge of Chaos

The "edge of chaos" is a heuristic for problem selection that can help to locate projects that might lead to long tails. The edge of chaos is used to refer to the space between a more ordered area and a chaotic area. Operating at the edge of chaos means wrangling a chaotic area and transforming a piece of it into something ordered, and this can produce very high returns.

There are many examples of the edge of chaos as a general phenomenon. In human learning, the zone of proximal development represents a level of difficulty (e.g. in school assignments) that is not so hard as to be incomprehensible, but not so easy as to require little thought. When building cellular automata, you need to take care to ensure the simulation is not so chaotic as to be incomprehensible but not so ordered as to be completely static. There's a narrow sweet spot where emergent, qualitatively distinct outcomes are possible. This is the area where it is possible for individuals to be a creative, highly impactful force.

In the context of safety research, staying on the edge of chaos means avoiding total chaos and total order. In areas with total chaos, there may be no tractability, and solutions are almost impossible to come by. This includes much of the work on "futuristic" risks: exactly which systems the risks will arise from is unclear, leading to a constant feeling of being unable to grasp the main problems. In the previous post, we argued that futuristic thinking is useful to begin to define problems, but for progress to be made, some degree of order must be made out of this chaos. However, in areas with total order, there is unlikely to be much movement since the low-hanging fruit has already been plucked.

Designing metrics is a good example of something that is on the edge of chaos. Before a metric is devised, it is difficult to make progress in an area or even know if progress has been made. After the development of a metric, the area becomes much more ordered and progress can be more easily made. This kind of conversion allows for a great deal of steering of resources towards an area (whatever area the new metric emphasizes) and allows for tail impact.

Another way to more easily access the edge of chaos is to keep a list of projects and ideas that don't work now, but might work later, for instance, after a change in the research field or an increase in capabilities. Periodically checking this list to see if any of the conditions are now met can be useful, since these areas are most likely to be near the edge of chaos. In venture capital, a general heuristic is to "figure out what can emerge now that couldn't before."

One useful edge of chaos heuristic is to only do one or two non-standard things in any given project. If a project deviates too much from existing norms, it may not be understood; but if it is too similar, it will not be original. At the same time, heavily imitating previous successes or what made a person previously successful leads to repetition, and risks not generating new value.

The following questions are also useful for determining if an area is on the edge of chaos: Have there been substantial developments in the area in the past year? Has thinking or characterization of the problem changed at all recently? Is it not obvious which method changes will succeed and which will fail? Is there a new paradigm or coherent area that has not been explored much yet (contrast with pre-paradigmatic areas that have been highly confused for a long time, which are more likely to be highly chaotic than at the edge of chaos)? Has anyone gotten close to making something work, but not quite succeeded?

We will now discuss specific high-leverage points for influencing AI safety. We note that they can be analogized to many of the processes discussed above.

Managing Moments of Peril

My intuition is that if we minimize the number of precarious situations, we can get by with virtually any set of technologies.

—Tyler Cowen

It is not necessary to believe this statement to believe the underlying implication: moments of peril are likely to precipitate the most existentially-risky situations. In common risk analysis frameworks, catastrophes arise not primarily from failures of components, but from the system overall moving into unsafe conditions. When tensions are running high or progress is moving extremely quickly, actors may be more willing to take more risks.

In cases like this, people will also be more likely to apply AI towards explicitly dangerous aims such as building weapons. In addition, in an adversarial environment, incentives to build power-seeking AI agents may be even higher than usual. As Ord writes:

Recall that nuclear weapons were developed during the Second World War, and their destructive power was amplified significantly during the Cold War, with the invention of the hydrogen bomb. History suggests that wars on such a scale prompt humanity to delve into the darkest corners of technology.

Better forecasting could help with either prevention or anticipation of moments of peril. Predictability of a situation is also likely to reduce the risk factor of humans making poor decisions in the heat of the moment. Other approaches to reducing the risk of international conflict are likely to help.

Because of the risks of moments of peril, we should be ready for them. During periods of instability, systems are more likely to rapidly change, which could be extremely dangerous, but perhaps also useful if we can survive it. Suppose a crisis causes the world to "wake up" to the dangers of AI. As Milton Friedman remarked: "Only a crisis – actual or perceived – produces real change. When that crisis occurs, the actions that are taken depend on the ideas that are lying around." A salient example can be seen with the COVID-19 pandemic and mRNA vaccines. We should make sure that the safety ideas lying around are as simple and time-tested as possible when a crisis will inevitably happen.

Getting in early

Building in safety early is very useful. In a report for the Department of Defense, Frola and Miller observe that approximately 75% of the most critical decisions that determine a system's safety occur early in development. The Internet was initially designed as an academic tool with neither safety nor security in mind. Decades of security patches later, security measures are still incomplete and increasingly complex. A similar reason for starting safety work now is that relying on experts to test safety solutions is not enough—solutions must also be time-tested. The test of time is needed even in the most rigorous of disciplines. A century before the four color theorem was proved, Kempe's peer-reviewed proof went unchallenged for years until, finally, a flaw was uncovered. Beginning the research process early allows for more prudent design and more rigorous testing. Since nothing can be done both hastily and prudently, postponing machine learning safety research increases the likelihood of accidents. (This paragraph is based on a paragraph from Unsolved Problems in ML Safety.)

As Ord writes, “early action is best for tasks that require a large number of successive stages.” Research problems, including ML problems, contain many successive stages. AI safety has and will also require a large number of successive stages to be successful: detecting that there’s a problem, clarifying the problem, measuring the problem, creating initial solutions, testing and refining those solutions, adjusting the formulation of the problem, etc. This is why we cannot wait until AGI to start to address problems in real ML systems.

Another reason for getting in early is that things compound: research will influence other research, which in turn influences other research, which can help self-reinforcing processes produce outsized effects. Historically, this has been almost all progress in deep learning. Such self-reinforcing processes can also be seen as an instance of preferential attachment.

Stable trends (e.g. scaling laws) lead people to question whether work on a problem will make any difference. For example, benchmark trends are *sometimes* stable (see the previous post for progress across time). However, it is precisely because of continuous research effort that new directions for continuing trends are discovered (cf. Moore's law). Additionally, starting/accelerating the trend for a safety metric earlier rather than later would produce clear counterfactual impact.

Scaling laws

Many different capabilities have scaling laws, and the same is true for some safety metrics. One objective of AI safety research should be to improve scaling laws of safety relative to capabilities.

For new problems or new approaches, naive scaling is often not the best way to improve performance. In these early stages, researchers with ideas are crucial drivers, and ideas can help to change both the slope and intercept of scaling laws.

To take an example from ML, consider the application of Transformers to vision. iGPT was far too compute-intensive, and researchers spent over a year making it more computationally efficient. This didn’t stand the test of time. Shortly thereafter, Google Brain, which is more ideas-oriented, introduced the “patchify” idea, which made Transformers for vision computationally feasible and resulted in better performance. The efficiency for vision Transformers has been far better than for iGPT, allowing further scaling progress to be made since then.

To take another example, that of AlphaGo, the main performance gains didn’t come from increasing compute. Ideas helped drive it forward (from Wikipedia):

Versions	Hardware	Elo rating	Date	Results
AlphaGo Fan	176 GPUs,[53] distributed	3,144[52]	Oct 2015	5:0 against Fan Hui
AlphaGo Lee	48 TPUs,[53] distributed	3,739[52]	Mar 2016	4:1 against Lee Sedol
AlphaGo Master	4 TPUs,[53] single machine	4,858[52]	May 2017	60:0 against professional players; Future of Go Summit
AlphaGo Zero (40 block)	4 TPUs,[53] single machine	5,185[52]	Oct 2017	100:0 against AlphaGo Lee 89:11 against AlphaGo Master
AlphaZero (20 block)	4 TPUs, single machine	5,018 [63]	Dec 2017	60:40 against AlphaGo Zero (20 block)

One can improve scaling laws by improving their slope or intercept. It's not easy to change the slope or intercept, but investing in multiple people who could potentially produce such breakthroughs has been useful.

In addition, for safety metrics, we need to move as far along the scaling law as possible, which requires researchers and sustained effort. It is usually necessary to apply exponential effort to continue to make progress in scaling laws, which requires continually increasing resources. As ever, social factors and willingness of executives to spend on safety will be critical in the long term. This is why we must prioritize the social aspects of safety, not just the technical aspects.

Scaling laws can be influenced by ideas. Ideas can change the slope (e.g., the type of supervision) and intercept (e.g., numerous architectural changes). Ideas can change the data resources: the speed of creating examples (e.g., saliency maps for creating adversarial examples), cleverly repurposing data from the Internet (e.g., using an existing subreddit to collect task-specific data), recognizing sources of superhuman supervision (such as those from a collective intelligence, such as a paper recommender based on multiple peoples' choices). Ideas can change the compute resources, for example through software-level and hardware-level optimizations improvements. Ideas can define new tasks and identify which scaling laws are valuable to improve.

Don't let the perfect be the enemy of the good

Advanced AI systems will not be ideal in all respects. Nothing is perfect. Likewise, high-risk technologies will be forced into conditions that are not their ideal operating conditions. Perfection in the real world is unattainable, and attempts to achieve perfection may not only fail, but they also might achieve less than attempts carefully aimed at reducing errors as much as possible.

For example, not all nuclear power plants meltdown; this does not mean there are no errors in those plants. *Normal Accidents* looked at organizational causes of errors and notes that some "accidents are inevitable and are, in fact, normal." Rather than completely eliminate all errors, the goal should be to minimize the impact of errors or prevent errors from escalating and carrying existential consequences. To do this, we will need fast feedback loops, prototyping, and experimentation. Due to emergence and unknown unknowns, risk in complex systems cannot be completely eliminated or managed in one fell swoop, but it can be progressively reduced. All else being equal, going from 99.9% safe to 99.99% safe is highly valuable. Across time, we can

continually drive up these reliability rates, which will continually increase our expected civilizational lifespan.

Sometimes it's argued that any errors at all with a method will necessarily mean that x-risk has not really been reduced, because an optimizer will necessarily exploit the errors. While this is a valid concern, it should not be automatically assumed. The next section will explain why.

Problems with asymptotic reasoning

In some parts of the AI safety community, there is an implicit or explicit drive for asymptotic reasoning or thinking in the limit. "Why should we worry about improving [safety capability] now since performance of future systems will be high?" "If we let [variable] be infinite, then wouldn't [safety problem] be completely solved?" "Won't [proposed safety measure] completely fail since we can assume the adversary is infinitely powerful?" While this approach arises from some good intuitions and has useful properties, it should not always be taken to the extreme.

Goodhart's Law

Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.

—Goodhart's Law (original phrasing, not the simplistic phrasing)

Goodhart's Law is an important phenomenon that is crucial to understand when conducting AI safety research. It is relevant to proxy gaming, benchmark design, and adversarial environments in general. However, it is sometimes misinterpreted, so we seek to explain our view of the importance of Goodhart's Law and what it does and does not imply about AI safety.

Goodhart's law is sometimes used to argue that optimizing a single measure is doomed to create a catastrophe as the measure being optimized ceases to be a good measure. This is a far stronger formulation than originally stated. While we must absolutely be aware of the tendency of metrics to collapse, we should also avoid falling into the trap of thinking that *all objectives can never change and will always collapse in all circumstances*. Strong enough formulations are tantamount to claiming that there is no goal or way to direct a strong AI safely (implying our evitable doom). Goodhart's Law does not prove this: instead, it shows that adaptive counteracting systems will be needed to prevent the collapse of what is being optimized. It also shows that metrics will not always include everything that we care about, which suggests we should try to include a variety of different possible goods in an AGI's objective. Whether we like it or not, all objectives are wrong, but some are useful.

Counteracting forces

There are many examples of organizations optimizing metrics while simultaneously being reeled in by larger systems or other actors from the worst excesses. For instance, while large businesses sometimes employ unsavory practices in pursuit of profits, in many societies they do not hire hitmen to assassinate the leaders of competing companies. This is because another system (the government) understands that the maximization of profits can create negative incentives, and it actively intervenes to prevent the worst case outcomes with laws.

To give another example, the design of the United States constitution was explicitly based on the idea that all actors would be personally ambitious. Checks and balances were devised to attempt to subdue the power of any one individual and promote the general welfare (as James Madison wrote, “ambition must be made to counteract ambition”). While this system does not always work, it has successfully avoided vesting all power in the single most capable individual.

Intelligence clearly makes a difference in the ability to enact counter forces to Goodhart’s Law. An extremely intelligent system will be able to subvert far more defenses than a less intelligent one, and we should not expect to be able to restrain a system far more intelligent than all others. This suggests instead that it is extremely important to avoid a situation where there is only a single agent with orders of magnitude more intelligence or power than all others: in other words, there should not be a large asymmetry in our offensive and defensive capabilities. It also suggests that the design of counteracting incentives of multiple systems will be critical.

In order to claim that countervailing systems are not appropriate for combating Goodhart’s Law, one may need to claim that offensive capabilities must always be greater than defensive capabilities, or alternatively, that the offensive and defensive systems will necessarily collude.

In general, we do not believe there is a decisive reason to expect offensive capabilities to be leagues better than defensive capabilities: the examples from human systems above show that offensive capabilities do not always completely overwhelm defensive capabilities (even when the systems are intelligent and powerful), in part due to increasingly better monitoring. We can’t take the offensive ability to the limit without taking the defensive ability to the limit. Collusion is a more serious concern, and must be dealt with when developing counteracting forces. In designing incentives and mechanisms for various countervailing AI systems, we must decrease the probability of collusion as much as possible, for instance, through AI honesty efforts.

Asymptotic reasoning recognizes that performance of future systems will be high, which is sometimes used to argue that work on counteracting systems is unnecessary in the long term. To see how this reasoning is overly simplistic, assume we have an offensive AI system, with its capabilities quantified with o, and a protective defensive AI system, with its capabilities quantified p. It may be true that o and p are high, but we also need to care about factors such as $p - o$ and the difference in derivatives $\frac{dp}{d\,\text{resources}} - \frac{do}{d\,\text{resources}}$. Some say that future systems will be highly capable, so we do not need to worry about improving their performance in any defensive dimension. Since the relative performance of systems matters and since the scaling laws for safety methods matter, asserting that all variables will be high enough not to worry about them is a low-resolution account of the long term.

Some examples of counteracting systems include artificial consciences, AI watchdogs, lie detectors, filters for power-seeking actions, and separate reward models.

Rules vs Standards

So, we’ve been trying to write tax law for 6,000 years. And yet, humans come up with loopholes and ways around the tax laws so that, for example, our multinational corporations are paying very little tax to most of the countries that they operate in. They find loopholes. And this is what, in the book, I call the loophole principle. It

doesn't matter how hard you try to put fences and rules around the behavior of the system. If it's more intelligent than you are, it finds a way to do what it wants.

—Stuart Russell

This is true because tax law is exclusively built on *rules*, which are clear, objective, and knowable beforehand. It is built on rules because the government needs to process hundreds of millions of tax returns per year, many tax returns are fairly simple, and people want to have predictability in their taxes. Because rule systems cannot possibly anticipate all loopholes, they are bound to be exploited by intelligent systems. Rules are fragile.

The law has another class of requirements, called standards, which are designed to address these issues and others. Standards frequently include terms like "reasonable," "intent," and "good faith," which we do not know how to assess in a mechanistic manner. We simply "know it when we see it:" in fact, a common legal term, *res ipsa loquitur*, means "the thing speaks for itself." Unlike rule-based code, deep neural networks can model these types of fuzzier concepts.

Unlike the tax code, which is based on rules and can be adjudicated by logic-based computer programs such as TurboTax, the criminal law is adjudicated by an intelligent system with intuitions (a judge and perhaps a jury). If a criminal is acquitted when they are guilty, it is because the intelligent system failed to collect enough evidence or interpret it correctly, not because the defense found a "loophole" in the definition of homicide (the exception is when lawyers make mistakes which create trouble under the *rules* used for procedure and evidence).

Russell's argument correctly concludes that rules alone cannot restrain an intelligent system. However, standards (e.g. "use common sense", "be reasonable") can restrain some intelligent behavior, provided the optimizing system is not too much more intelligent than the judiciary. This argument points to the need to have intelligent systems, rather than mechanistic rules, that are able to evaluate other intelligent systems. There are also defensive mechanisms that work for fuzzy raw data, such as *provable* adversarial robustness, that can help strengthen the defense. It is correct to conclude that an AGI's objectives should not be based around precise rules, but it does not follow that all objectives are similarly fragile.

Goal refinement

Goodhart's Law applies to *proxies* for what we care about, rather than what we actually care about. Consider ideal utilitarianism: does Goodhart's Law show that "maximizing the good" will inevitably lead to ruin? Regardless of how one views ideal utilitarianism, it would be wrong to conclude that it is refuted by Goodhart's Law, which warns that many *proxies* for good (e.g. "the number of humans who are smiling") will tend to collapse when subjected to optimization pressure.

Proxies that capture something we care about will likely have an approximation error. Some objectives have more approximation error than others: for instance, if we want to measure economic health, using real GDP reported by the US government will likely have less approximation error than nominal GDP reported in a text file on my computer. When subjected to optimization, that approximation error may become magnified, as optimizers can find areas where the approximation is particularly flawed and potentially manipulate it. This suggests that as optimization power increases, approximation error must correspondingly decrease, which can happen with better models, or approximation errors must become harder to exploit, which can happen with better

detectors. As such, systems will need to have their goals continuously refined and improved.

Methods for goal refinement might include better automated moral decision making and value clarification. We will discuss these in our next post.

Limitations of research based on a hypothetical superintelligence

Many research agendas start by assuming the existence of a superintelligence, and ask how to prove that it is completely safe. Rather than focus on microcosmic existing or soon-to-emerge systems, this line of research analyzes a model in the limit. This line of attack has limitations and should not be the only approach in the portfolio of safety research.

For one, it encourages work in areas which are far less tractable. While mathematical guarantees of safety would be the ideal outcome, there is good reason to believe that in the context of engineering sciences like deep learning, they will be very hard to come by (see the previous posts in the sequence). In information security, practitioners do not look for airtight guarantees of security, but instead try to increase security iteratively as much as possible. Even RSA, the centerpiece of internet encryption, is not provably completely unbreakable (perhaps a superintelligence could find a way to efficiently factor large numbers). Implicitly, the requirement of a proof and only considering worst-case behavior relies on incorrect ideas about Goodhart's Law: "if it is possible for something to be exploited, it certainly will be by a superintelligence." As detailed above, this account is overly simplistic and assumes a fixed, rule-based, or unintelligent target.

Second, the assumption of superintelligence eliminates an entire class of interventions which may be needed. It forces a lack of concretization, since it is not certain what kind of system will eventually be superintelligent. This means that feedback loops are extremely sparse, and it is difficult to tell whether any progress is being made. The approach often implicitly incentivizes retrofitting superintelligent systems with safety measures, rather than building safety into pre-superintelligent systems in earlier stages. From complex systems, we know that the crucial variables are often discovered by accident, and only empirical work is able to include the testing and tinkering needed to uncover those variables.

Third, this line of reasoning typically assumes that there will be a single superintelligent agent working directly against us humans. However, there may be multiple superintelligent agents that can rein in other rogue systems. In addition, there may be artificial agents that are above human level on only some dimensions (e.g., creating new chemical or biological weapons), but nonetheless, they could pose existential risks before a superintelligence is created.

Finally, asymptotically-driven research often ignores the effect of technical research on sociotechnical systems. For example, it does very little to improve safety culture among the empirical researchers who will build strong AI, which is a significant opportunity cost. It also is less valuable in cases of (not necessarily existential) crisis, just when policymakers will be looking for workable and time-tested solutions.

Assuming an omnipotent, omniscient superintelligence can be a useful exercise, but it should not be used as the basis for all research agendas.

Instead, improve cost/benefit variables

In science, problems are rarely solved in one fell swoop. Rather than asking, “does this solve every problem?” we should ask “does this make the current situation better?” Instead of trying to build a technical solution and then try to use it to cause a future AGI to swerve towards safety, we should begin steering towards safety now.

The military and information assurance communities, which are used to dealing with highly adversarial environments, do not search for solutions that render all failures an impossibility. Instead, they often take a cost-benefit analysis approach by aiming to increase the cost of the most pressing types of adversarial behavior. Consequently, a cost-benefit approach is a time-tested way to address powerful intelligent adversaries.

Even though no single factor completely guarantees safety, we can drive down risk through a combination of many safety features (defense in depth). Better adversarial robustness, ethical understanding, safety culture, anomaly detection, and so on to collectively make exploitation by adversaries harder, driving up costs.

In practice, the balance between the costs and benefits of adversarial behavior needs to be tilted in favor of the costs. While it would be nice to have the cost of adversarial behavior be infinite, in practice this is likely infeasible. Fortunately, we just need it to be sufficiently large.

In addition to driving up the cost of adversarial behavior, we should of course drive down the cost of safety features (an important high-level contributing factor). This means making safety features useful in more settings, easier to implement, more reliable, less computationally expensive, or have less steep or no tradeoffs with capabilities. Even if an improvement does not completely solve a safety problem once and for all, we should still aim to continue increasing the benefits. In this way, safety becomes something we can continuously improve, rather than an all-or-nothing binary property.

Some note we “only have one chance to get safety right,” so safety is binary. Of course, there are no do-overs if we’re extinct, so whether or not humans are extinct is indeed binary. However, we believe that the probability of extinction due to an event or deployment is not zero or one, but rather a continuous real value that we can reduce by cautiously changing the costs and benefits of hazardous behavior and safety measures, respectively. The goal should be to reduce risk as much as possible over time.

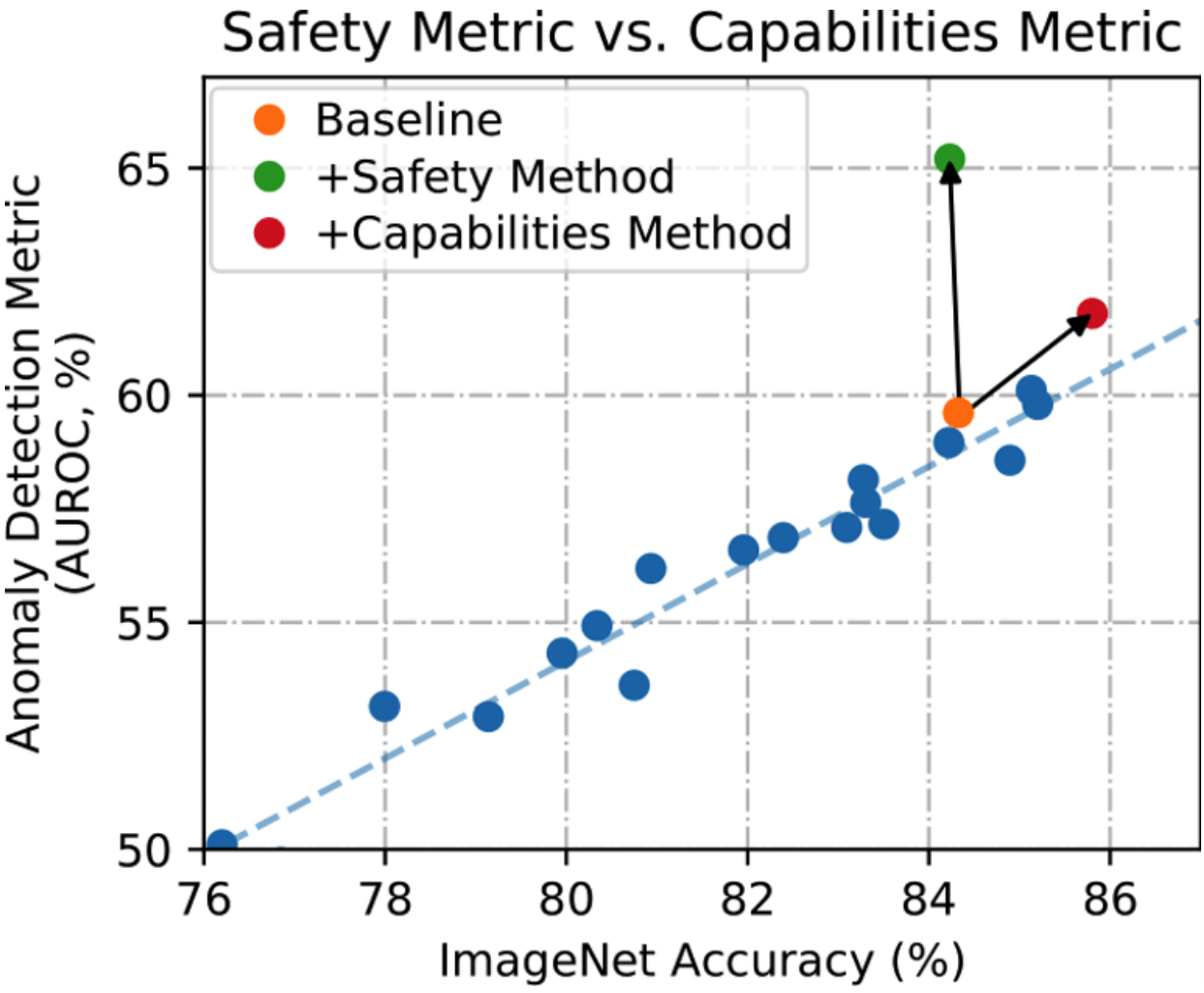
It’s important to note that not all research areas, including those with clear benefits, will have benefits worth their costs. We will discuss one especially important cost to be mindful of: hastening capabilities and the onset of x-risk.

Safety/capabilities tradeoffs

Safety and capabilities are linked and can be difficult to disentangle. A more capable system might be more able to understand what humans believe is harmful; it might also have more ability to cause harm. Intelligence cuts both ways. We do understand, however, that desirable behavior *can* be decoupled from intelligence. For example, it is well-known that *moral virtues* are distinct from *intellectual virtues*. An agent that is knowledgeable, inquisitive, quick-witted, and rigorous is not necessarily honest, just, power-averse, or kind.

In this section, by *capabilities* we mean *general capabilities.* These include general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities. We are not speaking of more specialized capabilities for downstream applications (for instance, climate modeling).

It is not wise to decrease some risks (e.g. improving a safety metric) by increasing other risks through advancing capabilities. In some cases, optimizing safety metrics might increase capabilities even if they aren't being aimed for, so there needs to be a more principled way to analyze risk. We must ensure that growing the safety field does not simply hasten the arrival of superintelligence.



The figure above shows the performance of various methods on standard ImageNet as well as their anomaly detection performance. The overall trendline shows that anomaly detection performance tends to improve along with more general ImageNet performance, suggesting that one way to make "safety progress" is simply to move along the trendline (see the red dot). However, if we want to make differential progress towards safety specifically, we should instead focus on safety methods that do not simply move along the existing trend (see the green dot). In addition, the trendline

also suggests that differential safety progress is in fact *necessary* to attain maximal anomaly detection performance, since even 100% accuracy would only lead to ~88% AUROC. Consequently researchers will need to shift the line up, not just move along the trendline. This isn't the whole picture. There may be other relevant axes, such as the ease of a method's implementation, its computational cost, its extensibility, and its data requirements. However, the leading question should be to ask what the effect of a safety intervention is on general capabilities.

It's worth noting that safety is commercially valuable: systems viewed as safe are more likely to be deployed. As a result, even improving safety without improving capabilities could hasten the onset of x-risks. However, this is a very small effect compared with the effect of directly working on capabilities. In addition, hypersensitivity to any onset of x-risk proves too much. One could claim that any discussion of x-risk at all draws more attention to AI, which could hasten AI investment and the onset of x-risks. While this may be true, it is not a good reason to give up on safety or keep it known to only a select few. We should be precautious but not self-defeating.

Examples of capabilities goals with safety externalities

Self-supervised learning and pretraining have been shown to improve various uncertainty and robustness metrics. However, the techniques were developed primarily for the purpose of advancing general capabilities. This shows that it is not necessary to be aiming for safety to improve it, and certain upstream capabilities improvements can simply improve safety "accidentally."

Improving world understanding helps models better anticipate consequences of their actions. It thus makes it less likely that they will produce unforeseen consequences or take irreversible actions. However, it also increases their power to influence the world, potentially increasing their ability to produce undesirable consequences.

Note that in some cases, even if research is done with a safety goal, it might be indistinguishable from research done with a capabilities goal if it simply moves along the existing trendlines.

Examples of safety goals with capabilities externalities

Encouraging models to be truthful, when defined as not asserting a lie, may be desired to ensure that models do not willfully mislead their users. However, this may increase capabilities, since it encourages models to have better understanding of the world. In fact, maximally truth-seeking models would be more than fact-checking bots; they would be general research bots, which would likely be used for capabilities research. Truthfulness roughly combines three different goals: accuracy (having correct beliefs about the world), calibration (reporting beliefs with appropriate confidence levels), and honesty (reporting beliefs as they are internally represented). Calibration and honesty are safety goals, while accuracy is clearly a capability goal. This example demonstrates that in some cases, less pure safety goals such as truth can be decomposed into goals that are more safety-relevant and those that are more capabilities-relevant.

One safety goal could be to incentivize collaboration, rather than competition, between different AI systems. This might be useful in reducing high-stakes conflicts that could lead to catastrophic outcomes. However, depending on how it is researched, it may come with capabilities externalities. For instance, focusing on getting agents to

perform better in positive-sum games might have a significant effect on general planning ability, which could have further downstream effects.

Better modeling “human preferences” may also be an example of a safety goal with capabilities externalities; we will cover this below.

Practical steps

When attempting to measure progress towards safety, it’s essential to also measure a method’s contribution to capabilities. One should ask whether a method creates a differential improvement in safety. Rather than relying on intuition to ascertain this, it is necessary to make empirical measurements. Empirical research claiming to differentially improve safety should demonstrate a differential safety improvement empirically. Of course, *reducing* capabilities is not likely to be helpful in practice, as this could make the method less likely to be used in the real-world.

Sometimes it is claimed that more general capabilities are needed to produce safety work, and so working on general capabilities advancements will at some point eventually allow working on safety. We agree that it is not necessarily the case that it could *never* be worth making capabilities advancements in exchange for differential improvements in safety. If at some point in the future it is impossible to make safety progress without an increase in capabilities, there may be more reason to accept capabilities externalities.

However, working on general capabilities for years to start studying a particular safety problem is neither precautious nor necessary. There are fortunately many safety research areas where it’s possible to make contributions without contributing to general capabilities at all. For instance, almost every paper in adversarial robustness hasn’t improved accuracy, because the two are not positively correlated. Similarly, out-of-distribution detection usually doesn’t come with capability externalities, and often focuses on eliciting out-of-distribution detection information from fixed models rather than improving their representations. We will discuss these and other areas and describe their relation to general capabilities in the next post.

An Application: Machine Ethics vs. Learning Task Preferences

Preference learning is typically operationalized as learning human preferences over different ways to accomplish a task. This is intended to ensure that agents understand what humans mean, rather than simply what they say. However, modeling “human values” or “human preferences” is often just modeling “user comparisons” or “task preferences,” not unlike the preference or comparison annotations that companies have been collecting for ML-driven translation, advertisement, and search algorithms throughout the past years. First, humans prefer smarter models. This is especially true when humans rate the usefulness of models. As such, modeling task preferences often does not pass the capability externalities test because it includes information about preferences for task-specific behavior (e.g. the quality of a summary). Second, preferences can be inconsistent, ill-conceived, and highly situation-dependent, so they may not be generalizable to the unfamiliar world that will likely arise after the advent of highly-capable models.

Consequently, we recommend trying to make models act in keeping with human values, not model preferences for a broad suite of general tasks. One area trying to do this is machine ethics, which is about building ethical AIs. (This is in contrast to AI ethics, which is about "ethics of AI" and is dominated by discussions of fairness, bias, and inequality; by way of its constituent's Foucaultian presuppositions, it often implicitly adopts anti-normative positions.) Rather than model task preferences, a core aim of machine ethics is modeling actual human values.

Compared with task preferences, ethical theories and human values such as intrinsic goods may be more generalizable, interpretable, and neglected. They are also more important to us (compared to preferences for high-quality summarization, for instance), and are also plausibly timeless. In addition, many normative factors are common to a number of ethical theories, even if theories disagree about how to combine them. Coarsely, normative factors are intrinsic goods, general constraints, special obligations, and options. An expansion of this list could be wellbeing, knowledge, the exercise of reason, autonomy, friendship, equality, culpability, impartiality, desert, deontological thresholds, intending harm, lying, promises, special obligations, conventions, duties to oneself, options, and so on. Note that these include factors that cover fairness, but also a whole spectrum of additional important factors.

In general, research into the application of ethical theories and the approximation of normative factors appears far less likely to lead to capabilities externalities, because the scope of what is being learned is restricted dramatically. Ethical theories contain less information that is relevant to understanding how to perform general tasks than generic human annotations and comparisons. Still, it's important to anticipate potential capabilities externalities: for example, one should not try to model consequentialist ethics by building better general predictive world models, as this is likely to create capabilities externalities.

One possible goal of machine ethics is work towards a moral parliament, a framework for making ethical decisions under moral and empirical uncertainty. Agents could submit their decisions to the internal moral parliament, which would incorporate the ethical beliefs of multiple stakeholders in informing decisions about which actions should be taken. Using a moral parliament could reduce the probability that we are leaving out important normative factors by focusing on only one moral theory, and the inherent multifaceted, redundant, ensembling nature of a moral parliament would also contribute to making the model less gameable. If a component of the moral parliament is uncertain about a judgment, it could request help from human stakeholders. The moral parliament might also be able to act more quickly to restrain rogue agents than a human could and act in the fast-moving world that is likely to be induced by more capable AI. We don't believe the moral parliament would solve all problems, and more philosophical and technical work will be needed to make it work, but it is a useful goal for the next few years.

Sometimes it is assumed that a sufficiently intelligent system will simply understand ethics, so there is no need to work on machine ethics. This analysis succumbs to the problems with asymptotic reasoning and assuming omniscience detailed above. In particular, we should not assume that an ethics model can automatically withstand the optimization pressure of another superintelligence, or that it will generalize in the same way as humans under distributional shift. We need to ensure that we will have aligned, reliable, and robust ethical understanding. A proactive ethics strategy is far more likely to succeed than one that naively hopes that the problem can be ignored or taken care of at the last moment. Additionally, on the sociotechnical front, people need time-tested examples if they are to be adopted or required in regulation. A moral parliament

will take years to engineer and accrue buy-in, so we cannot trust that our values will be best furthered by a last-minute few-shot moral parliament.

Conclusion

Starting research with asymptotic reasoning, while it has the benefit of aiming for research that has immediately graspable AI x-risk relevance, carries the cost of making research less specific and less tractable. It also reduces the number of research feedback loops.

By focusing on microcosms, empirical research is relevant for reducing AI x-risk, but its relevance is less immediately graspable. However, the reduction in immediately graspable relevance is more than made up for by increased tractability, specificity, measurability, and the information gained from faster feedback loops. Despite these strengths, naive empirical research threatens to produce capabilities externalities, which should be avoided as much as possible.

We propose a strategy to produce tractable tail impacts with minimal capabilities externalities. In summary:

- Pursue tail impacts, reduce moments of peril, start working on safety early, and improve the scaling laws of safety in comparison to capabilities.
- Since impact is likely to be tail distributed, it's important to understand where tail outcomes emerge from: multiplicative processes, preferential attachment, and the edge of chaos.
- "How can this safety mechanism make strong AI completely safe?" excludes many useful risk reduction strategies. Works that stand up the question "how can this work steer the AI development process in a safer direction?" are also useful for AI x-risk reduction.
- It's useful to view safety as a continuously improvable property rather than an all-or-nothing binary property.
- We take a stand against capabilities externalities in some safety research directions. AI safety research should be safe.
- Machine ethics should be preferred to learning task preferences, because the latter can have significant capability externalities, and ethics contains more time-tested and reliable values than task-specific preferences do.
- We suggest trying to achieve safety through evolution, rather than only trying to arrive at safety through intelligent design.

Open Problems in AI X-Risk [PAIS #5]

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

This is the fifth post in a sequence of posts that describe our models for Pragmatic AI Safety.

Dan Hendrycks (an author of this post), Nicholas Carlini, John Schulman, and Jacob Steinhardt previously wrote Unsolved Problems in ML Safety (2021), which lays out some of the most promising areas of research in ML safety. The paper is written for an academic audience; for the reasons discussed in previous posts, much of this audience would not have been receptive to a full discussion of existential risk. This post will present many of the same areas and a few more from the perspective of existential risk mitigation.

While some of the areas are well known in the AI safety community (honest AI), and others are well known in the broader ML community (such as adversarial robustness), many remain extremely neglected (such as power-averseness and moral decision-making). We hope to explain why these areas are relevant to existential risk.

The post will be presented mainly as a list of topics. The following table gives an overview of the problems with their importance, neglectedness, and tractability. This is not intended to be a list of all problems in AI safety, as we are focused on problems that are amenable to empirical ML research and where it is possible to avoid capabilities externalities.

Area	Problem	Importance	Neglectedness	Tractability
Alignment	Power-averseness	● ● ●	● ● ●	● ●
	Honest AI	● ● ●	● ● ●	● ●
	Moral Decision-Making	● ● ●	● ● ●	● ●
	Automated Moral Philosophy Research	● ● ●	● ● ●	●
Robustness	Adversarial Robustness	● ● ●	●	● ●
Monitoring	Anomaly Detection	● ● ●	● ●	● ●
	Interpretable Uncertainty	● ●	●	● ●
	Transparency	● ● ●	●	●
	Trojans	● ● ●	● ●	● ●
Systemic Safety	ML for Cyberdefense	● ●	● ● ●	● ● ●
	ML for Improving Epistemics	● ●	● ● ●	● ●
	Cooperative AI	● ● ●	● ● ●	●

Each area in this document will have the following sections:

1. Problem Description: a brief description of the problem
2. Motivation: an explanation of how the problem is relevant to AI x-risk. We include many different contributing motivations to each problem, and we do not believe that every individual motivation provides a decisive argument for the given area. Rather, the motivations together provide a good case for working on each of the areas.

3. What Researchers Are Doing Now: a list and explanation of prior work in the area.
4. What Advanced Research Could Look Like: a high-level overview of work that would make significant progress into this area. We mostly describe work that is likely to not be completed for at least a year or two.
5. Importance, Neglectedness, Tractability: a brief explanation of the ratings shown above.
6. Relation To General Capabilities: This section answers the question, “how much would progress in general capabilities help with this problem?”
7. Capabilities Externalities Analysis: This section answers the opposite question of the previous section, “how much would progress in this problem help with general capabilities?” As detailed in our previous post, capabilities externalities should be minimal for research areas at scale.
8. Criticisms: This section covers reasons that people give to argue an area is not valuable. We don’t necessarily agree with all of the critiques, and present them mainly for epistemic humility and so readers are familiar with some arguments against these problem areas.

Specific research project ideas for many of the areas are covered in Unsolved Problems in ML Safety, but we refrain from including these ideas in this document for brevity.

Alignment

We take Alignment to be about reducing inherent model hazards: hazards that result from models (explicitly or operationally) pursuing the wrong goals. Four concrete empirical research directions include honest AI, power-averseness, moral decision-making, and automated moral philosophy research. There are additional problems in alignment, but many have yet to be concretized.

Power-averseness

Problem Description

This area is about incentivizing models to avoid gaining more power than is necessary.

Motivation

Strategic AIs tasked with accomplishing goals would have instrumental incentives to accrue and maintain power, as power helps agents more easily achieve their goals. Likewise, some humans would have incentives to build and deploy systems that acquire power, because such systems would be more useful. If power-seeking models are misaligned, they could permanently disempower humanity.

If agents are given the power to single-handedly destroy humanity, a single system failure could result in an existential catastrophe. If power is instead distributed among multiple agents, failure could be decorrelated. This is more likely if agents are not constantly trying to overpower the others.

See Joe Carlsmith’s report for a more thorough motivation.

What Researchers Are Doing Now

We are currently working on developing power penalties, power limits, and taxonomizing and estimating model power. There has been some study on power-

seeking in general and the instrumental tendency to resist being shut off, but otherwise no attempt at power-averseness.

What Advanced Research Could Look Like

Models could evaluate the power of other agents in the world to accurately identify particular systems that were attaining more power than necessary. They could also be used to directly apply a penalty to models so that they are disincentivized from seeking power. Before agents pursue a task, other models could predict the types of power and amount of power they require. Lastly, models might be developed which are intrinsically averse to seeking power despite the instrumental incentive to seek power.

Importance, Neglectedness, Tractability

Importance: •••

This could reduce many inherent hazards. Models that do not accrue too much power would be easier to shut down and correct, and they could cause less damage.

Neglectedness: •••

Right now it's just a handful of people working on this.

Tractability: ••

There is an abundance of low-hanging fruit on the technical front, since almost no work has been performed in this area. However, this area may be less tractable because it relies on sequential decision making, which is not as developed as other areas of machine learning.

Relation to General Capabilities

Power-averseness is likely harmed by upstream improvements. As power-seeking becomes a more viable strategy for goal achievement, it may be increasingly incentivized instrumentally.

Capabilities Externalities Analysis

Power-averseness would harm general capabilities by default, as it would restrict the options available to the model somewhat. While it may improve the safety–capabilities balance, we will need to continually work towards making power-averseness techniques increasingly robust to competition and productivity pressures. In any case, capabilities externalities are avoided by default.

Criticisms

This could make models less economically valuable and especially less valuable during war, which is a large obstacle. Power-averseness will face many challenges on the sociotechnical front. What military wants this functionality, unless there is international coordination? What entity wants to limit its power?

It is impossible to overcome or counteract the instrumentally convergent drive for power, so efforts to do this will inevitably fail.

Honest AI

Problem Description

Honest AI involves creating models that only output what they hold to be true. It also involves determining what models hold to be true, perhaps by analyzing their internal representations.[1]

Motivation

If it is within a model's capacity to be strategically deceptive (i.e. able to make statements that the model in some sense knows to be false in order to gain an advantage) then treacherous turn scenarios are more feasible. Models could deceive humans about their plans, and then execute them once humans are no longer able to course-correct. Plans for a treacherous turn could be brought to light by detecting dishonesty, or models could be made inherently honest, allowing operators to query them about their true plans.

Other motivation formulations:

- We would like to prevent models from producing deceptive information.
- If models can be made honest and only assert what they believe, then they can produce outputs that are more representative and give human monitors a more accurate impression of their beliefs.
- Honesty helps facilitate cooperation among AIs, so it enables possibilities in cooperative AI. Honesty also undercuts collusion.

What Researchers Are Doing Now

They are demonstrating that models can lie, and they are capturing true and false clusters inside models (this paper is forthcoming).

What Advanced Research Could Look Like

Good techniques could be able to reliably detect when a model's representations are at odds with its outputs. Models could also be trained to avoid dishonesty and allow humans to correctly conclude that models are being honest with high levels of certainty.

Importance, Neglectedness, Tractability

Importance: •••

This could reduce many inherent hazards. If models were completely honest, deception would be far more difficult, thereby greatly reducing the probability of a whole class of failure modes.

Neglectedness: •••

This is a new area, though the idea of “faithful outputs” is a very easy sell to the ML community, so its neglectedness will probably decrease soon.

Tractability: ••

Research is in its early stages, but there is some initial (forthcoming) research that makes progress.

Relation to General Capabilities

There is not much evidence that honesty improves with general capabilities by default.

In fact, dishonesty may become more of a viable strategy for models with more ability to succeed at it.

Capabilities Externalities Analysis

Honesty is a narrower concept than truthfulness and is deliberately chosen to avoid capabilities externalities, since truthful AI is usually a combination of vanilla accuracy, calibration, and honesty goals. Optimizing vanilla accuracy is optimizing general capabilities, and we cover calibration elsewhere. When working towards honesty rather than truthfulness, it is much easier to avoid capabilities externalities.

Lie detection could uncover hidden knowledge in models, which could potentially allow inducing more advanced functionality. Techniques that focus on a model's concept of truth rather than querying its knowledge might help avoid this.

Criticisms

Current works use contrived situations and goad models to produce lies rather than studying lies produced under realistic conditions. The current honesty tools are very fragile.

Adaptive, self-aware models in the future could circumvent current honesty techniques.

If we select heavily for models that are powerful and honest, we might select for models that do not know that their general tendency is to gain the upper hand but which still do have that tendency (models for which deceptiveness becomes an unknown known rather than a known known). In other words, power-seeking might become "unconscious" rather than "conscious."

Honesty could make models very unpleasant since "the truth is terrible". An extremely honest model might become a cynic activist that tears down everyone's fundamental beliefs, so full honesty may be neither economically attractive nor compatible with human psychological safety.

Implementing Moral Decision-Making

Problem Description

This area is about building models to understand ethical systems and steering models to behave ethically.

This research area includes a few strategies:

- Model intrinsic goods and normative factors, as these will be relevant even under extreme world changes. This is in contrast to task preferences; book summarization preferences are less fundamental human values, and their relevance is more fragile under extreme world changes.
- Given moral systems, get models to abide by them with an artificial conscience or other forms of endogenous self-regulation.
- Implement an automated moral parliament to have models act appropriately in the face of moral uncertainty.

A generalization of this area is machine ethics, which is about making models that act ethically; this is an alternative formulation of Alignment.

Motivation

This line of work helps create actionable ethical objectives for systems to pursue. If strong AIs are given objectives that are poorly specified, they could pursue undesirable actions and behave unethically. If these strong AIs are sufficiently powerful, these misspecifications could create an existential catastrophe. Consequently, work in this direction helps us avoid proxy misspecification as well as value lock-in.

If our foremost goal is reducing the probability of destroying or permanently curtailing the potential of humanity, then it seems to make most sense to focus on aligning AI to the most important and time-tested values, namely those considered in normative ethics.

Other potential motivations:

- Robustness is easier to achieve in a limited area than it is to achieve across a very wide range of tasks. If there's one place we want to ensure robustness, it's moral decision-making.
- An artificial conscience can block morally suspect actions in AI systems by having direct access to action choices.

What Researchers Are Doing Now

They are predicting when there is high moral disagreement for a scenario. They are modeling normative factors and intrinsic goods, implementing foundational ethical theories, modeling the provenance of utility functions, modeling exceptions to moral rules, researching how to more effectively steer artificial agents' actions (such as through an artificial conscience), and so on.

What Advanced Research Could Look Like

High-functioning models should detect situations where the moral principles apply, assess how to apply the moral principles, evaluate the moral worth of candidate actions, select and carry out actions appropriate for the context, monitor the success or failure of the actions, and adjust responses accordingly.

Models could represent various purported intrinsic goods, including pleasure, autonomy, the exercise of reason, knowledge, friendship, love, and so on. Models should be able to distinguish between subtly different levels of these goods, and these value functions should not be vulnerable to optimizers. Models should be able to create pros and cons of actions with respect to each of these values, and brainstorm how changes to a given situation would increase or decrease the amount of a given intrinsic good. They should also be able to create superhuman forecasts of how an action can affect these values in the long-term (e.g., how studying can reduce wellbeing in the short-term but be useful for wellbeing in the long-term), though this kind of research must be wary of capabilities externalities. Models should also be able to represent more than just intrinsic goods, as they should also be able to represent constantly-updating legal systems and normative factors including special obligations and deontological constraints.

Another possible goal is to create an automated moral parliament, a framework for making ethical decisions under moral and empirical uncertainty. Agents could submit

their decisions to the internal moral parliament, which would incorporate the ethical beliefs of multiple stakeholders in informing decisions about which actions should be taken. Using a moral parliament could reduce the probability that we are leaving out important normative factors by focusing on only one moral theory, and the inherent multifaceted, redundant, ensembling nature of a moral parliament would also contribute to making the model less gameable. If a component of the moral parliament is uncertain about a judgment, it could request help from human stakeholders. The moral parliament might also be able to act more quickly to restrain rogue agents than a human could and act in the fast-moving world that is likely to be induced by more capable AI. We don't believe the moral parliament would solve all problems, and more philosophical and technical work will be needed to make it work, but it is a useful goal for the next few years.

Importance, Neglectedness, Tractability

Importance: •••

Models that make decisions with regard for their morality would be far less likely to cause catastrophes.

Neglectedness: •••

A handful of people are working on this.

Tractability: ••

So far, there has been continual progress.

Relation to General Capabilities

Moral decision-making can benefit from upstream capabilities. Models that are better able to understand the world in general will be more able to understand how morality fits into that world. Better predictive power would also enable better modeling of consequentialist moral theories.

Capabilities Externalities Analysis

This has similarities with task preference learning ("I like this Netflix movie"; "I like this summary more"), but the latter has obvious externalities: humans prefer smarter models. Instead, we try to model normative factors and intrinsic goods. Task preferences are less robust and relevant under extreme environmental changes, compared to enduring human values such as normative factors (wellbeing, impartiality, etc.) and the factors that make up a good life (pursuing projects, gaining knowledge, etc.). Capabilities externalities are readily avoidable, provided that one is not modeling task preferences.

We should model consequentialist theories using pre-existing general world model capabilities, rather than try to build better predictive world models for the sake of modeling consequentialist theories. Doing so will keep capabilities externalities to a minimum.

Criticisms

In order to do groundbreaking work, it is useful for the researcher to have taken a course or two in normative ethics. Few have, which makes this area less accessible. In

addition, many researchers weak in normative ethics might attempt research and produce low-quality yet influential work.

Models should learn to do what humans would do, not abide by abstract moral theories. Morality does not model what humans do well, and we should care about how humans really behave rather than how they think they should behave.

“Alignment” should not be associated with ethics because that will intimidate technical researchers; this is why we should talk about task preferences.

Ethics is not yet resolved enough to leave anything up to a machine. It would thus be better to model preferences (implicitly, preference utilitarianism) rather than complicated explicit moral theories.

Developing systems that perform moral decision making could reduce the influence of humans in making ethical decisions, reducing our autonomy.

Automated Moral Philosophy Research (Value Clarification)

Problem Description

This area is about building AI systems that can perform moral philosophy research. This research area should utilize existing capabilities and avoid advancing general research, truth-finding, or contemplation capabilities.

Motivation

The future will sharpen and force us to confront unsolved ethical questions about our values and objectives. In recent decades, peoples’ values have evolved by confronting philosophical questions, including whether to infect volunteers for science, how to equitably distribute vaccines, the rights of people with different orientations, and so on. How are we to act if many humans spend most of their time chatting with compelling bots and not much time with humans, and how are we to balance pleasure and enfeeblement? Determining the right action is not strictly scientific in scope, and we will need philosophical analysis to help us correct structural faults in our proxies.

To address deficiencies in our moral systems, and to more rapidly and wisely address future moral quandaries that humanity will face, these research systems could help us reduce risks of value lock-in by improving our moral precedents earlier rather than later. If humanity does not (or cannot) take a “long reflection” to consider and refine its values after it develops strong AI, then the value systems lying around may be amplified and propagated into the future. Value clarification reduces risks from locked-in, deficient value systems. Additionally, value clarification can be understood as a way to reduce proxy misspecification, as it can allow values to be updated in light of new situations.

We will need to decide what values to pursue and how to pursue them. If we decide poorly, we may lock in or destroy what is of value. It is also possible that there is an ongoing moral catastrophe, which we would not want to replicate across the cosmos.

What Researchers Are Doing Now

Nothing.

What Advanced Research Could Look Like

Good work in value clarification would be able to produce original insights in philosophy, such that models could make philosophical arguments or write seminal philosophy papers. Value clarification systems could also point out inconsistencies in existing ethical views, arguments, or systems.

Importance: •••

This would reduce many ontological or systematic misalignment errors.

Neglectedness: •••

No one is working on this.

Tractability: •

This is a challenging problem. (Possible intermediate steps are described in Unsolved Problems.)

Relation to General Capabilities

General upstream research capabilities will make this problem more tractable.

Philosophical/fuzzy reasoning ability and raw intelligence seem distinct; by default high-IQ educated people are not especially good at reasoning about fuzzy abstract objects.

Capabilities Externalities Analysis

We do not aim to make models superhuman at research generally, only superhuman at moral philosophy. We turn to moral philosophy research rather than general research or general-purpose wide reflective equilibrium approximators since moral philosophy research can have few capabilities externalities, while the latter proposals have extremely high superhuman capabilities externalities. Poorly directed research in this direction could have high capabilities externalities, but it is likely this problem will remain neglected relative to general truth-seeking/contemplation/reflection methods, so the best way to improve automated moral philosophy research using the marginal researcher is to use existing capabilities and apply them to this problem.

Criticisms

There is no progress in moral philosophy. Alternatively, perhaps moral philosophy does not provide any useful insights about what the world ought to look like or what agents ought to do. If normative ethics has provided no insight, it will not be useful to model it.

Intelligence will straightforwardly result in expert philosophical ability, so we do not need to specifically work on it.

We can just wait for the long reflection to correct structural forms of misalignment, assuming we get to the long reflection. The danger of lock-in is overrated.

Nobody will trust works of philosophy that include genuine moral progress if they are generated by a machine.

Robustness

We take Robustness to be about reducing vulnerabilities to hazards from sources other than the model itself.

Robustness focuses on responding to abnormal, unforeseen, unusual, highly impactful, or adversarial events. Adversarial robustness does not currently have economically feasible scaling laws, even for extremely simplistic adversaries. Even on MNIST, an extremely simple dataset, no model has human-level robustness, and we haven't been able to get any perception task that requires machine learning to a human level of robustness.

Robustness is not the same thing as in-distribution accuracy: even if a model gets higher in-distribution accuracy, it does not necessarily provide high robustness. If you rely on more of the existing trend, you do not get reliability. Tesla is trying to improve robustness with petabytes of task-specific data, and yet the problem remains unsolved.

Adversarial Robustness

Problem Description

Adversarial examples demonstrate that optimizers can easily manipulate vulnerabilities in AI systems and cause them to make egregious mistakes. Adversarial vulnerabilities are long-standing weaknesses of AI models. While typical adversarial robustness is related to AI x-risk, future threat models will be broader than today's adversarial threat models. Since we are concerned about being robust to optimizers that cause models to make mistakes generally, we should make minimal assumptions about the properties of the adversary and work to make models that are robust to many kinds of attacks.

Motivation

In the future, AI systems may pursue goals specified by other AI proxies. For example, an AI could encode a proxy for human values, and another AI system could be tasked with optimizing the score assigned by this proxy. If the human value proxy is not robust to optimizers, then its vulnerabilities could be exploited, so this gameable proxy may not be fully safe to optimize. By improving the reliability of learned human value proxies, optimizers and adversaries would have a harder time gaming these systems. If gaming becomes sufficiently difficult, the optimizer can be impelled to optimize the objective correctly. Separately, humans and systems will monitor for destructive behavior, and these monitoring systems need to be robust to adversaries.

We often study adversarial robustness in the continuous domain (vision) because gradient descent is powerful in that setting. In contrast, adversarial attacks for text-based models are weaker and often not gradient-based. We'd like to defend against the most powerful adversaries since as models get more intelligent, the attacks will get more powerful. Gradient attacks for vision systems are already extremely strong, so it gives us a more direct view of how to defend against powerful optimizers.

Other motivation formulations:

- Adversarial robustness is helpful for safety community building efforts and improving safety culture.
- This reduces risk by reducing vulnerability to hazards. (Risk_Hazard = Hazard x Exposure x Vulnerability)

- Adversarial robustness is necessary for extreme and worst-case reliability.

What Researchers Are Doing Now

- Producing new distortions of given images [1,2,3] and text [1,2] as adversarial examples.
- Finding ways to robustify models (adversarial training improvements) [1,2,3,4,5,6]
- Adversarially constructed datasets [1,2]

Importance, Neglectedness, Tractability

Importance: •••

High reliability is necessary for safe superintelligence. Work on this problem is a necessary component of high reliability.

Neglectedness: •

Numerous researchers are working in this area. However, aspects are neglected. Adversarial robustness for larger-scale systems is rarely studied due to academic researchers lacking the necessary compute. Likewise, long-term threat models are underexplored: attack specifications may not be known beforehand, and attack budgets could be large.

Tractability: ••

There are shovel-ready problems and the research is (slowly) progressing.

What Advanced Research Could Look Like

Ideally, an adversarially robust system would make reliable decisions given adversarially constructed inputs, and it would be robust to adversaries with large attack budgets using unexpected novel attacks. Furthermore, it should detect adversarial behavior and adversarially optimized inputs. A hypothetical human value function should be as adversarially robust as possible so that it becomes safer to optimize. A hypothetical human value function that is fully adversarially robust should be safe to optimize.

Relation to General Capabilities

Adversarial robustness is barely helped by upstream capabilities.

Capabilities Externalities Analysis

For vision, nearly all methods in adversarial robustness do not improve general capabilities such as clean accuracy, so in this problem area it is quite easy to avoid capabilities externalities. Furthermore, adversarial training for vision currently dramatically reduces vanilla classification accuracy.

In text, automated or virtual adversarial attacks don’t markedly improve general capabilities much. However, human-crafted natural adversarial examples for text models seem to improve general capabilities.

Criticisms

Most research uses particular types of perturbations on images and the main field ignores more important things (unforeseen attacks, large distortions).

Currently, there does not appear to be that much knowledge transfer between continuous signal (vision) and discrete sequence (text) adversarial robustness research. It's unclear how big of a role continuous signals will play for strong AI outside perception.

Current adversarial robustness methods have very large general capabilities costs. Currently it's not orthogonal to general capabilities–it's anticorrelated. While researchers are iteratively reducing the cost, it could still be a costly safety measure in the future.

The instant two intelligent agents can reason about each other–regardless of their goals–they will necessarily collude. Adversarial robustness, in trying to improve the defensive capabilities of proxy models against other models, will not be relevant if the models collude.

Monitoring

We take monitoring to mean avoiding exposure to hazards as much as possible, such as by detecting problems in systems before they grow worse.

Anomaly Detection

Problem Description

This area is about detecting potential novel hazards such as unknown unknowns, unexpected rare events, or emergent phenomena. Anomaly detection (also known as out-of-distribution detection) can allow models to flag salient anomalies for human review or execute a conservative fallback policy.

Motivation

This is an indispensable tool for detecting a wide range of hazards. For example:

- proxy gaming
- rogue ML systems that are already causing harm at smaller scales
- deceptive ML systems not easily detectable by humans
- Trojan horse models (discussed below)
- malicious users who may attempt to intentionally misalign a model, or align it to their own nefarious ends
- early signs of dangerous novel technologies
- AI tripwires could help uncover early misaligned systems before they can cause damage
- emergent behavior

This approach operates by virtue of being able to detect other hazards before they occur or before they can cause more damage. While in an ideal world the other hazards do not occur at all, in reality any serious attempt at safety must build in some mechanisms to detect failures to prevent hazards. Early detection is crucial to being able to successfully stop the hazard before it becomes impossible to do so.

In addition to helping with AI x-risk, anomaly detection can also be used to help detect novel engineered microorganisms that present biological x-risks, perhaps by having image and sequence anomaly detectors scan hospitals for novel pathogens (see this paper for an example of anomaly detection involving genomic sequences). Anomaly detection could also help detect Black Balls ("a technology that invariably or by default destroys the civilization that invents it").

Other motivation formulations:

- This reduces risk by reducing exposure to hazards. (Risk_Hazard = Hazard x Exposure x Vulnerability)
- Detection has a central place in Hazard Analysis (e.g., Failure Modes and Effects Analysis). It is customary for improving defense in depth.
- This can help flag examples, models, or trajectories for human review.
- This helps systems fail gracefully by enabling us to implement triggers for fail-safes or conservative fallback policies. This reduces the probability of hard maximizing models going off the rails.
- This can be used to detect malicious use (malicious use that uses novel sly tactics are harder to detect). As before, malicious use could be from AIs or humans.
- Various other systems have this as a main line of defense (high reliability organizations, information security operations centers, the human body, etc.) and it has a central place in Hazard Analysis by creating defense in depth. Specifically, it reduces risk by reducing exposure to hazards. (Risk_Hazard = Hazard x Exposure x Vulnerability). Anomaly detection could also be used to detect malicious use, whether by AIs or humans.
- An approach towards safety is building conservative agents that refrain from taking unprecedented actions. Unprecedented actions could be blocked by anomaly detectors.

What Researchers Are Doing Now

- Out-of-distribution detection aims to identify when new input data is out of the original training distribution, and flag this for possible review [1,2,3,4,5,6].
- One-class learning aims to use data from a single distribution to train a classifier to detect when new data is out-of-distribution [1,2]
- More applied subproblems: For text processing, researchers detect if genetic instructions belong to a new species. In the future, researchers can try to detect whether a utility function is applicable to a given text input (does this scenario describe a sentient being or something else)? In vision, researchers detect if an image contains an organism species that the model has not seen before (this includes microorganisms).

What Advanced Research Could Look Like

A successful anomaly detector could serve as an AI watchdog that could reliably detect and triage rogue AI threats. When the watchdog detects rogue AI agents, it should do so with substantial lead time. Anomaly detectors should also be able to straightforwardly create tripwires for AIs that are not yet considered safe. Furthermore, advanced anomaly detectors should be able to help detect black balls. Anomaly detectors should also be able to detect biological hazards by increasing detection lead time by having anomaly detectors continually scan hospitals for novel biological hazards.

Importance, Neglectedness, Tractability

Importance: •••

This cross-cuts many relevant problems.

Neglectedness: ••

After many years of community building, there are many researchers working in this area. The most cited paper in the area only has somewhat more than 1,000 citations, whereas there is an adversarial robustness paper with more than 10,000 citations.

Adversarial anomaly detection is more neglected. Anomaly detection for text-based models is also more neglected.

Tractability: ••

Detecting image anomalies is one of the hardest problems in image understanding, along with adversarial robustness. Progress is even slower than in adversarial robustness, though that may be because the field is smaller.

Relation to General Capabilities

It's somewhat helped by upstream capabilities, but extrapolating accuracy to 100% does not come close to solving anomaly detection even in simple, non-adversarial settings. It is harmed by other common upstream methods (e.g., mixup).

Capabilities Externalities Analysis

Most research in anomaly detection takes fixed models and tries to extract better anomaly scores. When models are fixed, capabilities externalities are largely avoided. Other research changes the representation of models to make them more suitable for anomaly detection; these methods are not that useful for making models more generally capable.

In this setting, externalities are especially easy to avoid.

Criticisms

Detecting a rogue model at the point when it's virtually unstoppable will not help us stop it. By the time something is going wrong, we are already dead.

This does not directly reduce inherent model hazards; it just detects them: this could be used to detect AI deception, but it would not directly make the system desire to be honest. Sometimes this is just a protective measure and not a preventative measure. For making other high risk technologies safer, protective measures are less desirable than preventative measures.

Along with general progress in AI, this increases the feasibility of totalitarianism at the cost of better responding to hazards.

When models are more intelligent, anomalies will be easy to detect. (However, it's quite difficult to argue that even a superintelligence could detect all Black Swans the moment they start to emerge and all malicious actors employing a novel strategy and that detection lead time cannot be increased.)

Researchers are only slightly focused on adversarial anomaly detection.

Interpretable Uncertainty

Problem Description

This area is about making model uncertainty more interpretable and calibrated by adding features such as confidence interval outputs, conditional probabilistic predictions specified with sentences, posterior calibration methods, and so on.

Motivation

If operators ignore system uncertainties since the uncertainties cannot be relied upon or interpreted, then this would be a contributing factor that makes the overall system that monitors and operates AIs more hazardous. To draw a comparison to chemical plants, improving uncertainty calibration could be similar to ensuring that chemical system dials are calibrated. If dials are uncalibrated, humans may ignore the dials and thereby ignore warning signs, which increases the probability of accidents and catastrophe.

Furthermore, since many questions in normative ethics have yet to be resolved, human value proxies should incorporate moral uncertainty. If AI human values proxies have appropriate uncertainty, there is a reduced risk in an human value optimizer maximizing towards ends of dubious value.

Other reasons:

- Calibrated models can better convey the limits of their competency by expressing their uncertainty, so human operators can know when to override models. Calibrated probabilities facilitate rational decision making:
    - Improved probability estimates matter for high-stakes decisions
    - Improved risk estimates (probabilities multiplied by losses)
- ML subsystems are easier to integrate if each system is well-calibrated
- Model confidences are more interpretable the more they are calibrated
- This helps systems fail gracefully by enabling us to implement triggers for fail-safes or conservative fallback policies.

What Researchers Are Doing Now

They are measuring model miscalibration on typical examples and in the face of distribution shifts and adversarial examples [1,2,3,4,5].

What Advanced Research Looks Like

Future models should be calibrated on inherently uncertain, chaotic, or computationally prohibitive questions that extend beyond existing human knowledge. Their uncertainty should be easily understood by humans. Moreover, given a lack of certainty in any one moral theory, AI models should accurately and interpretably represent this uncertainty in human value proxies.

Importance, Neglectedness, Tractability

Importance: ••

This is an important part of interpretability.

Neglectedness: •

Many people are working on it, maybe half an order of magnitude more than anomaly detection. Calibration in the face of adversaries is highly neglected, as are new forms of interpretable uncertainty: having models output confidence intervals, having models output structured probabilistic models (e.g., “event A will occur with 60% probability assuming event B also occurs, and with 25% probability if event B does not”).

Tractability: ••

There are shovel-ready tasks, and the community is making progress on this problem.

Relation to General Capabilities

It’s often helped by some upstream capabilities, but it is harmed by other upstream methods (e.g., mixup). Also note the Brier score metric is much more correlated with upstream capabilities metrics such as vanilla accuracy than calibration metrics including the expected calibration error (ECE) or RMS calibration error. See a discussion of how the Brier score is a mixture of an accuracy component and a calibration component in this EMNLP paper. Consequently, safety-minded researchers should avoid tangling under- and over-confidence with accuracy and therefore avoid using the Brier score as the main summary for calibration. Fortunately most of the ML community uses the disentangled metrics.

Capabilities Externalities Analysis

Many calibration methods try to make fixed models more calibrated, and these techniques leave the representations and accuracy unchanged. By default, calibration research leaves capabilities unchanged.

Criticisms

Like work in transparency, this helps human operators and inspectors, but it does not directly reduce inherent hazards.

Many of the impacts are indirect or sociotechnical, but we should only support work that has direct impact (linear causal influence).

Trojan Horse Models

Problem Description

AI systems can contain “trojan” hazards. Trojaned models behave typically in most situations, but when specific secret situations are met, they reliably misbehave. For example, an AI agent could behave normally, but when given a special secret instruction, it could execute a coherent and destructive sequence of actions. In short, this area is about identifying hidden functionality embedded in models that could precipitate a treacherous turn.

Motivation

One of the most dangerous sources of risk from advanced AI is sudden, unexpected changes in behavior. Similar to how people can hide their true intentions, a misaligned AI could bypass oversight mechanisms through deception. If we can uncover hidden behavior and predict treacherous turns before they happen, this will mitigate several failure modes.

What Researchers Are Doing Now

They are developing Trojan attacks and defenses. Most existing work uses CV datasets and models as a testbed [1,2,3,4,5,6,7], but recent work is beginning to explore Trojans for NLP models [1,2]. A much smaller number of papers explores Trojans for RL [1]. There is also related work on emergent behaviors in RL and emergent capabilities in large language models, which explores different aspects of hidden functionality.

Importance, Neglectedness, Tractability

Importance: •••

Treacherous turns from advanced AI systems are a significant source of x-risk. Starting work on this problem early is important, and Trojan research is one way to make initial progress.

Neglectedness: ••

The field of Trojans in deep learning is 5 years old, but there is still much left to be done. The field is not commonly associated with safety, so there is an opportunity to focus the field towards greater x-risk relevance and create a path for safety researchers to gain career capital.

Tractability: ••

Analyzing and detecting Trojans is an early field with much low-hanging fruit. It is a standard ML research problem with emerging benchmarks that can be iterated on. However, the problems of detecting and reverse-engineering Trojans are broad in scope and challenging.

Relation to General Capabilities

We are not aware of more accurate models affecting Trojan attacks or defenses.

Capabilities Externalities Analysis

Trojan research is unlikely to impact general capabilities, because most of the work is developing attacks and defenses. The former work is highly specific to Trojans and unlikely to transfer, and the latter work is relevant to safety.

Conversely, improvements to general capabilities might yield useful demonstrations of hidden functionality and emergent capabilities that could make this research easier. However, this doesn't appear to have happened yet.

Criticisms

Trojans might not generalize to real treacherous turns. Current Trojans are mostly about flipping predicted classes, but real hidden behavior might be more complex. Even future harder versions of the problem may not solve treacherous turns.

Real hidden behavior might be easy to predict, e.g. malintent can simply be read off from the AI's train of thought, particularly if progress has been made on honesty. Large-scale deception will be hard for strong AIs to maintain.

Real hidden behavior might be hidden very differently from human-created Trojans, and so Trojan detection methods might not help.

Transparency

Problem Description

AI systems are becoming more complex and opaque. This area is about gaining clarity about the inner workings of AI models and making models more understandable to humans.

Motivation

If humans lose the ability to meaningfully understand ML systems, they may no longer retain their sovereignty over model decisions.

Transparency tools could help unearth deception, mitigating risks from dishonest AI and treacherous turns. This is because some speculate that deception could become inadvertently incentivized, and if models are capable planners, they may be skilled at obscuring their deception. Similarly, researchers could develop transparency tools to detect poisoned models, models with trojans, or models with other latent unexpected functionality. Moreover, transparency tools could help us better understand strong AI systems, which could help us more knowledgeably direct them and anticipate their failure modes.

Other motivation formulations:

- Transparency helps facilitate cooperation among AIs, so it enables possibilities in cooperative AI. Transparency also undercuts collusion.
- Transparency could make it easier for models to detect deception and other problems in other models, which could improve monitoring.

What Researchers Are Doing Now

People are critiquing transparency methods, analyzing superhuman game AIs, and looking for mechanisms inside models. This is quite a simplification. There are many researchers in this space, but there aren't many coherent clusters of research (save for areas known to be bad, such as nearly all work on saliency maps).

What Advanced Research Could Look Like

Successful transparency tools would allow a human to predict how a model would behave in various situations without testing it. These tools should be able to be easily applied (ex ante and ex post emergence) to unearth deception, emergent capabilities, and failure modes.

To help make models more transparent, future work could try to provide clarity about the inner workings of models and understanding model decisions. Another line of valuable work is critiquing explainability methods and trying to show limitations of auditing methods. Measuring similarities and differences between internal representations is also an important step toward understanding models and their latent representations.

Importance, Neglectedness, Tractability

Importance: •••

If we could intuitively understand what models are doing, then they'd be far more controllable.

Neglectedness: •

This is highly funded by numerous stakeholders, and it has a large community. Deep nets are famous for being "black boxes," and this limits their economic utility due to concerns about human oversight (such as in medical applications).

Tractability: •

This area has been struggling to find a solid line of attack throughout its existence. It has set goals for itself, and it has not met them (e.g., using transparency tools to find special functionality implanted by another human.)

Relation to General Capabilities

The historical trend is that models are becoming more opaque as time progresses. We went from decision trees and SVMs, to random forests (intellectually unmanageable), to ConvNets, to Vision Transformers (where the lack of channels makes feature visualizations worse). "In prediction, accuracy and simplicity (interpretability) are in conflict" (Breiman, 2001).

Capabilities Externalities Analysis

This has not been useful for addressing weaknesses in models, and the "insights" gleaned from transparency techniques have not helped researchers understand how to improve models. It currently is not having any impact on capabilities research.

(It is worth noting that if "transparency" is construed more broadly to include anything involving "understanding models better," then some approaches may have capabilities externalities, such as studying general capability scaling laws. We do not include this as a transparency problem.)

Criticisms

There's been high interest from numerous stakeholders, including several tens of millions of dollars from DARPA, and there isn't anything solid that we've got from it. Maybe it's too vague or trying to do too much?

If we care about detecting deception or Trojan horse models (treacherous turns), why not just focus on the more tractable problem of detecting deception or Trojan horse models directly (possibly using AIs to help us detect such behavior and sift through the gobs of raw data)?

Progress in ML is driven by metrics; progress must be measurable. This area doesn't have metrics, while other areas such as detecting Trojans do.

Interpretability is a "god of the gaps" field ("if we understood our models, then we could understand how to fix this problem."). For many non-accuracy goals that we don't know how to reach, many think interpretability will help us reach the goal. (The Mythos of Interpretability)

Often, systematic findings that explain a portion of a model's behavior are "science of ML" papers (example) rather than intrpretability papers.

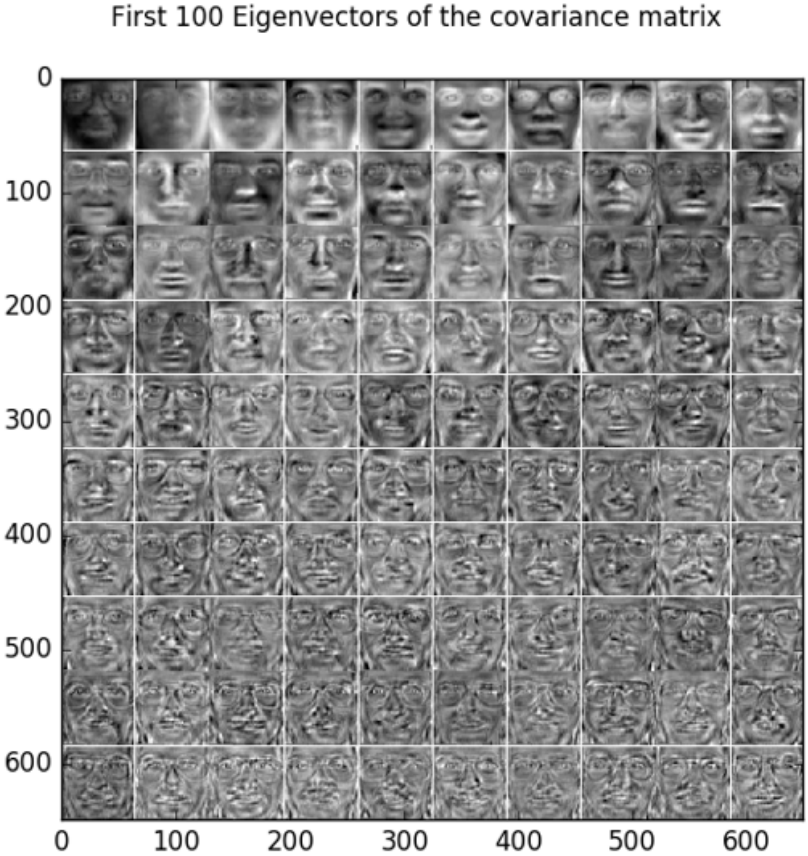
Transparency aims to reverse the longstanding trend that models are becoming more opaque; reversing or undoing the consequences of this robust trend is unlikely to happen.

(Even in humans, better forecasting models are less interpretable and more complicated.)

Humans are not interpretable. They are black boxes to themselves, and they confabulate explanations. (Hinton gives an example.) Expecting explanations for thought processes requires making many unconscious computations conscious, but only a small fraction of computation can be supported consciously. Many intuitive tasks are pre-verbal.

If humans could in detail explain how they processed an image and classified it, then we could close the book on several research areas in cognitive science. Likewise, if humans could understand how they know things, then research experts could easily transmit their research abilities. We know they're relatively unable to do this, or else committed students could always become stronger than advisors. ("Education is an admirable thing," wrote Oscar Wilde, "but it is well to remember from time to time that nothing that is worth knowing can be taught." Moreover, "we know more than we can tell." - Michael Polanyi)

It is difficult to interpret the eigenvectors from PCA of simple datasets such as MNIST or faces:



It’s not possible to easily convert the vectors above to English. How should we expect to be able to interpret even more complicated ML algorithms if it’s not possible to interpret PCA? Likewise, other basic ML methods are hard to interpret. For example, random forests are intellectually unmanageable. The later computations in a basic MNIST ConvNet are hardly interpretable.

There are often too many possible explanations for a phenomenon for any single explanation to be useful. This is why the best test of a model is its predictive power, not whether it can only explain prior data. NYU professor Bob Rehder wrote: “Explanation tends to induce learners to search for general patterns, it may cause them to overlook exceptions, with the result that explanation may be detrimental in domains where exceptions are common.”

When analyzing complex systems (such as deep networks), it is tempting to separate the system into events or components (“parts”), analyze those parts separately, and combine results or “divide and conquer.”

This approach often wrongly assumes (Leveson 2020):

- Separation does not distort the system’s properties
- Each part operates independently
- Part acts the same when examined singly as when acting in the whole
- Parts are not subject to feedback loops and nonlinear interactions
- Interactions between parts can be examined pairwise

Searching for mechanisms and reductionist analysis is too simplistic when dealing with complex systems (see our third post for more).

People hardly understand complex systems. Grad students in ML don’t even understand various aspects of their field, how to make a difference in it, what trends are emerging, or even what’s going on outside their small area. How will we understand an intelligence that moves more quickly and has more breadth? The reach of a human mind has limits. Perhaps a person could understand a small aspect of an agent’s actions (or components), but it’d be committing the composition fallacy to suggest a group of people that individually understand a part of an agent could understand the whole agent.

There are different approaches to understanding phenomena. One can look from the top down and find general rules that capture complex phenomena. Alternatively, one could try building understanding from the bottom up, but building a staircase from basic mechanisms all the way to complex system-level behavior does not have a good track record (e.g., consider mechanisms in evolution that leave many adaptations unexplained, mechanisms in anatomy have limited power in predicting whether a new drug will work, etc.). Most current transparency builds from the bottom up.

Smarter people are not reliably understood by less smart people. If a less intelligent person could reliably predict what a more intelligent person would do, then they would be similarly intelligent. It’s unlikely we’ll be able to understand all aspects of a superintelligence (make it interpretable), as the human mind is limited and going to be less smart than a superintelligence.

Even if we labeled each of a model’s many millions of neurons (example label of a neuron: “detects whiskers at 21 to 24 degrees, and it also detects black letter ‘A’ keyboard keys with probability 37%, and it detects a type of noise that’s difficult to describe”), it wouldn’t necessarily be interpretable (simulable, enable crisp post-hoc

explanations, highly decomposable). To holistically understand models, we are better off understanding them functionally.

Systemic Safety

This section is about using AI for furthering longtermist or EA goals. The areas are also useful for improving systemic contributing factors that contribute to the reduction of AI x-risk.

ML for Cyberdefense

Problem Description

This area is about using machine learning to improve defensive security, such as by improving malicious program detectors. This area focuses on research avenues that are clearly defensive and not easily repurposed into offensive techniques, such as detectors and not automated pentesters.

Motivation

It will matter very little if AI systems are aligned if they can be hacked by humans or other AI systems and made to be misaligned (intentionally or unintentionally). There may also be situations where aligned AI is hijacked by a malicious actor who intentionally or accidentally contributes to x-risk. In addition, one of the fastest and most potent ways for a superintelligence to project its intelligence and influence the world is through cyberattacks, not through physical means.

Even if some of the components of ML systems are safe, they can become unsafe when traditional software vulnerabilities enable others to control their behavior. Moreover, traditional software vulnerabilities may lead to the proliferation of powerful advanced models, and this may be worse than proliferating nuclear weapons.

Cyberattacks could take down national infrastructure including power grids, and large-scale, reliable, and automated cyberattacks could engender political turbulence and great power conflicts. Great power conflicts incentivize countries to search the darkest corners of technology to develop devastating weapons. This increases the probability of weaponized AI, power-seeking AI, and AI facilitating the development of other unprecedented weapons, all of which are x-risks. Using ML to improve defense systems by decreasing incentives for cyberwarfare makes these futures less likely.

Other motivation formulations:

- As ML proliferates, it's possible that we will see a dramatic increase in the use of ML systems in cyberattacks. In order to protect against such attacks, we may need to use ML in cyberdefense. Work might be especially necessary if there turns out to be an asymmetry in attacks/defenses as there currently is with adversarial examples.
- This post also gives motivations for the importance of information security in AI safety.

What Researchers Are Doing Now

Some research in automatically detecting and patching software vulnerabilities at scale [1,2,3]. Some use anomaly detection for use in detecting malicious payloads [1,2]. ML

has also been used for intrusion detection [1,2,3].

What Could Advanced Research Look Like?

AI-based security systems could be used for better intrusion detection, firewall design, malware detection, and so on.

Importance, Neglectedness, Tractability

Importance: ••

It's important that powerful systems do not fall into the hand of extreme or reckless actors, but they may be able to develop those systems themselves regardless.

Neglectedness: •••

Security researchers are currently bottlenecked by compute power.

Tractability: •••

This is an application of ML, and doesn't necessarily require new fundamental research.

Relation to General Capabilities

Better upstream models could help make for better ML defenses. For example, better reasoning and the ability to understand longer sequences could make models better able to analyze code and large assembly files.

Capabilities Externalities Analysis

Much of the work in this space is engineering and not influencing upstream models.

Criticisms

This should only be funded or researched in an exploratory way before large commitments are made.

Funding will need to strongly disincentivize gain-of-function attack capabilities in order for this area to be positive.

This is just realpolitik as it assumes that systems are much less likely to be safe if they fall into the wrong hands. In reality, systems are unlikely to be roughly equally safe (or unsafe) in any case, so devoting time to cybersecurity will not help.

If nuclear weapons aren't much of an x-risk, then WWIII is unlikely to be a large x-risk. It might be necessary to improve security for particular ML models to prevent them from falling into the wrong hands, but better security in general isn't useful.

ML for Improving Epistemics

Problem Description

This area is about using machine learning to improve the epistemics and decision-making of political leaders. This area is tentative; if it turns out to have difficult-to-avoid capabilities externalities, then it would be a less fruitful area for improving safety.

Motivation

We care about improving decision-making among political leaders to reduce the chance of rash or possibly catastrophic decisions. These decision-making systems could be used in high-stakes situations where decision-makers do not have much foresight, where passions are inflamed, and decisions must be made extremely quickly and based on gut decisions. During these moments of peril, humans are liable to make egregious errors. Historically, the closest we have come to a global catastrophe has been in these situations, including the Cuban Missile Crisis. Work on these technologies could reduce the prevalence of perilous situations. Separately, this reduces the risks from persuasive AI. Moreover, it helps leaders more prudently wield the immense power that future technology will provide. As Carl Sagan said, “If we continue to accumulate only power and not wisdom, we will surely destroy ourselves.”

Other motivation formulations:

- Better forecasting can potentially help with instituting better regulations and calibrating AI strategy. In addition, it could reduce risks from hasty deployments predicated on other actors being farther along than they are.
- This reduces x-risks from hyper-persuasive AI and an erosion of epistemics.
- AI can help political leaders make better decisions, like what we needed during emerging crises like COVID.

What Researchers Are Doing Now

We are developing ML benchmarks for forecasting geopolitical events.

What Advanced Research Could Look Like

Systems could eventually become superhuman forecasters of geopolitical events. They could help brainstorming possible considerations that might be crucial to a leader’s decision. Finally, they could help identify inconsistencies in a leader’s thinking and help them produce a more sound judgment.

Importance, Neglectedness, Tractability

Importance: ••

Better epistemics could be useful for the development and deployment of AI systems, but it would not solve any fundamental problems in AI safety on its own.

Neglectedness: •••

Few care about superforecasting, let alone ML for forecasting.

Tractability: ••

This is an application of ML, but it is fairly outside the capabilities of current models.

Relation to General Capabilities

Forecasting ability and IQ are not strongly related. Exceptional forecasting skills seem to be hard to acquire even for smart people. Better retrieval methods could help improve forecasting capabilities.

Capabilities Externalities Analysis

Much of the work in this space is engineering and not influencing upstream systems.

If work on this problem is appearing to play into general capabilities, work should be discouraged (this goes for any emerging safety research area). It is important to keep this line of research targeted to reduce the chances of speeding up other kinds of capabilities (for instance, truth/contemplation/reasoning/research).

Criticisms

“The essence of intelligence is prediction.” Therefore it may be harder to avoid capabilities externalities.

Taleb argues in *Antifragile* that reliance on forecasting makes us more vulnerable to tail risks and gives us a false sense of security.

Forecasting benchmarks need to span decades of historical data to measure the ability to predict tail risks. That will require a substantial engineering effort.

Cooperative AI

Problem Description

In the future, AIs will interact with humans and other AIs. For these interactions to be successful, models will need to be more skilled at cooperating. This area is about reducing the prevalence and severity of cooperation failures. AI models and humans may be stuck in poor equilibria that are robustly difficult to escape; cooperative AI methods should improve the probability of escaping or avoiding poor equilibria. This problem also works towards making AI agents better at positive-sum games, of course subject to capabilities externalities constraints. As we describe this area, it does not include the typical directions in human-robot interaction, such as communication between humans and robots in standard tasks.

Motivation

First, worlds where multiple agents are aligned in different ways are highly plausible. There are strong incentives to have multiple decision-making agents; for example, jury theorems show collections of agents make better decisions than a single agent, and agents have incentives to retain some control and not automatically cede control to one single centralized agent.

In a world where we have AIs interacting with other agents, cooperative AI can be useful for not just having higher upside but also smaller downside. Cooperative AIs could help rein in misaligned agents or power-seeking AIs. For this protective measure to work, the power of the collective must be greater than the power of the power-seeking AI.

Let’s consider how easily a power-seeking AI could overpower the world. Of course, if AIs are better able to cooperate, they are more likely to counteract power-seeking AIs. Tracking and regulating the flow and concentration of GPUs can reduce the probability of a single AI becoming more powerful than the collective power of the rest. Even if one power-seeking agent is smarter than every other model, it does not imply that it has control over all other models. Usually, having higher cognitive ability does not let an agent overpower the collective (the highest IQ person does not rule the world). However, individual bad actors that are smarter than others can have outsized effects. In some special environments, such as environments with structured criticality, small

differences could be magnified. Moreover, the world is becoming more long-tailed and more like “extremistan” in which there is tyranny of the top few (this is in contrast to mediocristan, where there is tyranny of the collective). Consequently, while there are factors that can give smarter models outsized advantages, the smartest model does not automatically overpower the collective power of cooperative AIs.

Other motivation formulations:

- To make environments with multiagent AIs stable, cooperation may be necessary.
- Cooperation can make us better able to escape bad equilibria, such as help us overcome bad systems for which we are dependent on our basic needs. Being robust to coordination failures helps us avoid lock-in.
- Intrasystem goals can create misalignment; cooperation helps agents better jointly optimize a decomposed objective, thereby reducing misalignment.
- If each superintelligence is fully aligned with one individual, the superintelligences and their owners can end up in game-theoretic tragedies
- The theory of morality as cooperation (“all of human morality is an attempt to solve a cooperative problem”) implies that if we want to build ethical machines (machine ethics), then it can be fruitful to make them more cooperative and build in cooperative dispositions.
- If we improve cooperativeness, we have a larger range of governance and deployment strategies (e.g., we could have AIs team up against other defecting AIs, this affects the strategic landscape and the possible defensive solutions), rather than trust a single contemplative agent. This is work towards making more sophisticated contractarian approaches to machine ethics more feasible (rather than only philosopher-king type strategies).
- Without cooperation, we will have hierarchies where the strongest agents dominate, leading to “the state of nature” and conflict; to help avoid such a dire environment, we need cooperation.
- Cooperation is a “defense in depth” area that does not decidedly fix safety problems, but it helps drive down the severity and probability of many hazards
- Cooperation could help rein in power-seeking or colluding AIs; a group of AIs, in many cases, have enough power to rein in misaligned AIs.
- Cooperation reduces the probability of conflict and makes the world less politically turbulent. Similarly, cooperation enables collective action to counteract rogue actors, regulate systems with misaligned goals, and rein in power-seeking or colluding AIs. Finally, cooperation reduces the probability of various forms of lock-in and helps us overcome and replace inadequate systems that we are dependent on for our basic needs.

What Advanced Research Could Look Like

Researchers could create agents that, in arbitrary real-world environments, exhibit cooperative dispositions (e.g., help strangers, reciprocate help, have intrinsic interest in others achieving their goals, etc.). Researchers could create coordination systems or AI agent reputation systems. Cooperating AIs should also be more effective at coordinating to rein in power-seeking AI agents.

Importance, Neglectedness, Tractability

Importance: •••

Within systemic safety, cooperative AI is the most targeted towards the reduction of AI x-risk.

Neglectedness: •••

There are few researchers working in this area.

Tractability: •

It is currently especially difficult to perform meaningful research on multiagent sequential decision making.

Relation to General Capabilities

Agents able to plan on very long time horizons may have more incentives to cooperate, but even for humans cooperative tendencies and precedents were hard-earned and hard to enforce; powerful humans often prefer not to cooperate but rather dominate.

Capabilities Externalities Analysis

Some research strategies could make agents better at playing games, which would make them better at playing cooperative games. By pursuing less naive research strategies, such as the strategy of endowing models with intrinsic dispositions to cooperate, capabilities externalities should be easier to avoid.

Research in this area should avoid developing cooperation methods that are antisymmetric with collusion and may therefore create collusive abilities. That is, research should avoid collusion externalities. Fortunately, some cooperative tendencies are not antisymmetric with collusion; being disposed to help strangers could be thought antisymmetric with the disposition to harm strangers, but the former helps cooperation and the latter does not help collusion. Cooperativeness is also useful for a larger, highly permeable group, while collusion is maintained among a specific group. Cooperation is often helped with transparency and truth, which is more robust than secrecy and lies, which collusion often depends on. Separately, working on honesty disincentivizes collusion.

Criticisms

Cooperation is too closely linked to collusion to be worth pursuing.

Cooperation leads to higher connectivity, which leads to higher fragility (faster cascading failures) and more extreme tail events (long tail distributions get sharper with more connectivity; then you have a more volatile future). Higher connectivity undermines the power of the collective, as these environments are dominated by tail events and the most extreme agents.

Possible additional areas

The section above represents concrete problems that we believe can be pursued now without capability externalities and with varying levels of tractability. In this section we discuss some possible additional areas. One of these areas is not concrete enough yet, and the other areas have not formed into a coherent area yet. These limitations may someday be resolved.

Regulating Mesa-Optimizers and Intrasystem Goals

As systems make objectives easier to optimize and break them down into new goals, subsystems are created that optimize these new intrasystem goals. But a common failure mode is that “intrasystem goals come first.” These goals can steer actions instead of the primary objective. Thus a system’s explicitly written objective is not necessarily the objective that the system operationally pursues, and this can result in misalignment.

Intrasystem goals occur when the goal of a training process (e.g., the loss function used for gradient descent, the exploration incentives of the sequential decision making agent, etc.) differs from the operational goal of the trained model it produces. This is known as mesa-optimization.

When multi-agent sequential decision-making is more feasible, we can give agents goals and delegate subgoals to agents. Since breaking down goals can distort them, this creates “intrasystem goals” and misalignment. Regulating these subagents that are optimizing their subgoal will be a research challenge. However, capabilities will need to be advanced further before this research area will be tractable.

An alternative way to study mesa optimizers is to study the general inductive biases of optimizers. While this could potentially be informative for understanding mesa-optimization, the neglectedness and tractability are low: this was previously the hottest area of theoretical ML for some years in the late 2010s, and see here for a discussion of tractability.

A related area that is more neglected and tractable is certified behavior. It is possible to have guarantees about model behavior given their weights [1,2], so it is not necessarily true that “all bets are off” when models are deployed.

Proxy Gaming

This is not clearly an area in its own right. Right now it looks like adversarial robustness, anomaly detection, and detecting emergent functionality, applied to sequential decision making problems. Perhaps in the future a distinct problem area will emerge.

Irreversibility

To avoid lock-in, some want to train models to pursue easy-to-reverse states. One way is to increase optionality; however, current methods to do this might simultaneously increase power-seeking behavior. Perhaps in the future avoiding irreversibility generally and preventing lock-in can be separated from power-seeking. Right now it seems there are other more targeted approaches to avoiding lock-in, such as moral parliaments, philosophy research bot/value clarification, and cooperative AI.

Conclusion

It is sometimes argued that the AI safety field has few specific problems that can be tractably pursued without creating capabilities externalities. However, we have shown that there are some specific research directions that can be pursued while avoiding capabilities externalities.

Many of the research directions listed in this document can be tractably pursued by the broader ML research community, making them suitable for broader outreach beyond the small group of researchers solely motivated by existential safety.

Though we believe some of these areas are more promising than others, we specifically do not argue for the overriding importance of a single one for reasons of diversification. We also do not claim that the areas above are the only areas worth pursuing, and like all research avenues, they may need to be curtailed in the future if they prove intractable or produce unacceptable externalities. We believe that the areas above are promising enough to be included in an overall scalable portfolio.

1. ^

   Note that detecting lies would best fit under Monitoring rather than Alignment, but for simplicity we consolidate these approaches here.