CS 690: Human-Centric Machine Learning

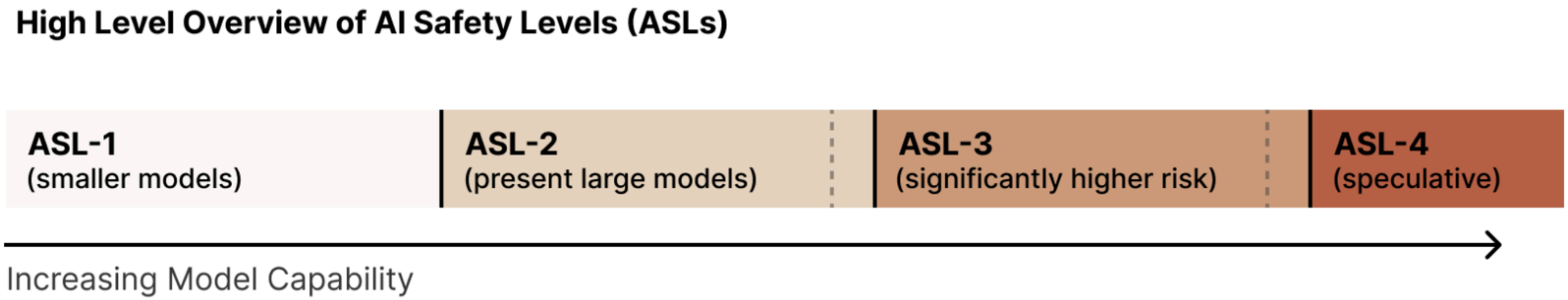
Prof. Scott Niekum

AI risk, mitigation, and counterarguments

Anthropic responsible scaling policy

Goal: Mitigate catastrophic risks — large-scale devastation (for example, thousands of deaths or hundreds of billions of dollars in damage) that is directly caused by an AI model and wouldn’t have occurred without it.

ASL levels



Anthropic's commitment to follow the ASL scheme thus implies that we commit to pause the scaling[2] and/or delay the deployment of new models whenever our scaling ability outstrips our ability to comply with the safety procedures for the corresponding ASL.

ASL risks

For each ASL, the framework considers two broad classes of risks:

- **Deployment risks:** Risks that arise from *active use* of powerful AI models. This includes harm caused by users querying an API or other public interface, as well as misuse by internal users (compromised or malicious). Our **deployment safety measures** are designed to address these risks by governing when we can safely deploy a powerful AI model.
- **Containment risks:** Risks that arise from merely *possessing* a powerful AI model. Examples include (1) building an AI model that, due to its general capabilities, could enable the production of weapons of mass destruction if stolen and used by a malicious actor, or (2) building a model which autonomously escapes during internal use. Our **containment measures** are designed to address these risks by governing when we can safely train or continue training a model.

ASL risks

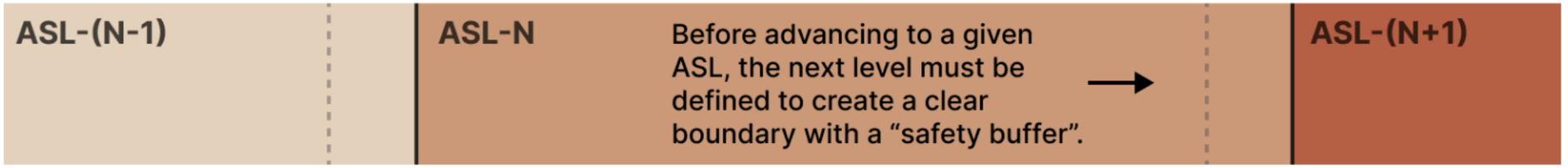
Sources of Catastrophic Risk

Our current understanding suggests at least two general sources of catastrophic risk from increasingly powerful AI models. For our initial commitments, we design our evaluations and safety measures with these risks in mind:

- **Misuse**: AI systems are dual-use technologies, and so as they become more powerful, there is an increasing risk that they will be used to intentionally cause large-scale harm, for example by helping individuals create CBRN[3] or cyber threats.
- **Autonomy and replication**: As AI systems continue to scale, they may become capable of increased autonomy that enables them to proliferate and, due to imperfections in current methods for steering such systems, potentially behave in ways contrary to the intent of their designers or users. Such systems could become a source of catastrophic risk even if no one deliberately intends to misuse them.

Iterative definitions

Rather than try to define all future ASLs and their safety measures now (which would almost certainly not stand the test of time), we will instead take an approach of *iterative* commitments. By iterative, we mean we will define ASL-2 (current system) and ASL-3 (next level of risk) now, and commit to define ASL-4 by the time we reach ASL-3, and so on.



ASL summary

AI Safety Level	Dangerous Capabilities	Containment Measures *Required to store model weights*	Deployment Measures *Required for internal/external use*
ASL-1	Models which *manifestly and obviously* pose no risk of catastrophe. For example, an LLM from 2018, or an AI system trained only to play chess.	None	None
ASL-2 *Our current safety level*	No capabilities likely to cause catastrophe, although early indications of these capabilities. For example, an AI system that can provide bioweapon-related information that couldn't be found via a search engine, but does so too unreliably to be useful in practice.	Evaluate for ASL-3 warning signs when training, using methods and *Evaluation Protocol* described below. Harden security against opportunistic attackers.	Follow current deployment best practices e.g. model cards, acceptable use policies, misuse escalation procedures, vulnerability reporting, harm refusal techniques, T&S tooling, and partner safety evaluation. These overlap significantly with our White House voluntary commitments.
ASL-3 *We are currently preparing these measures*	Low-level autonomous capabilities *or* Access to the model would substantially increase the risk of catastrophic misuse, either by proliferating capabilities, lowering costs, or enabling new methods of attack, as compared to a non-LLM baseline of risk.	Harden security such that non-state attackers are unlikely to be able to steal model weights and advanced threat actors (e.g. states) cannot steal them without significant expense. Evaluate for ASL-4 warning signs when training, likely similar to but much more involved than the methods described below. Implement internal compartmentalization for training techniques and model hyperparameters.	Implement strong misuse prevention measures, including internal usage controls, automated detection, a vulnerability disclosure process, and maximum jailbreak response times. Each deployed modality (e.g. API, fine-tuning) must pass intensive expert red-teaming and evaluation measures for catastrophic risks.
ASL-4	*Capabilities and warning sign evaluations defined before training ASL-3 models*		

ASL-2: Example deployment measures

While ASL-2 models do not carry significant risk of causing a catastrophe, their deployment still poses a range of trust and safety, legal, and ethical risks. To address these risks, our ASL-2 deployment commitments include:

- **Model cards:** Publish model cards for significant new models describing capabilities, limitations, evaluations, and intended use cases. The most recent model card for Claude 2 is available here.
- **Acceptable use:** Maintain and enforce an acceptable use policy (AUP) that restricts, at a minimum, catastrophic and high harm use cases, including using the model to generate content that could cause severe risks to the continued existence of humankind, or direct and severe harm to individuals. See our current AUP here which briefly describes our enforcement measures, which include maintaining the option to restrict access if extreme misuse issues emerge.
- **Vulnerability reporting:** Provide clearly indicated paths for our consumer and API products where users can report harmful or dangerous model outputs or use cases. Users of claude.ai can report issues directly in the product, and API users can report issues to usersafety@anthropic.com.
- **Harm refusal techniques**: Train models to refuse requests to aid in causing harm, such as with Constitutional AI or other improved techniques.
- **T&S tooling**: Require model enhanced trust and safety detection and enforcement. Claude.ai, our native API, and our distribution partners currently use a classifier model to identify harmful user prompts and model completions[6]. If automated fine-tuning is provided, data should similarly be filtered for harmfulness, and models should be subject to automated evaluation to ensure harmlessness features are not degraded.

ASL-3 Capabilities and Threat Models

We define an ASL-3 model as one that can either immediately, or with additional post-training techniques corresponding to less than 1% of the total training cost, do at least one of the following two things. (By post-training techniques we mean the best capabilities elicitation techniques we are aware of at the time, including but not limited to fine-tuning, scaffolding, tool use, and prompt engineering.)

1. **Capabilities that significantly increase risk of misuse catastrophe:** Access to the model would substantially increase the risk of deliberately-caused catastrophic harm, either by proliferating capabilities, lowering costs, or enabling new methods of attack. This increase in risk is measured relative to today's baseline level of risk that comes from e.g. access to search engines and textbooks. We expect that AI systems would first elevate this risk from use by **non-state attackers**[7].

   Our first area of effort is in evaluating bioweapons risks where we will determine threat models and capabilities in consultation with a number of world-class biosecurity experts. We are now developing evaluations for these risks in collaboration with external experts to meet ASL-3 commitments, which will be a more systematized version of our recent work on frontier red-teaming. In the near future, we anticipate working with CBRN, cyber, and related experts to develop threat models and evaluations in those areas before they present substantial risks. However, we acknowledge that these evaluations are fundamentally difficult, and there remain disagreements about threat models.

2. **Autonomous replication in the lab:** The model shows early signs of autonomous self-replication ability, as defined by 50% aggregate success rate on the tasks listed in [Appendix on Autonomy Evaluations]. The appendix includes an overview of our threat model for autonomous capabilities and a list of the basic capabilities necessary for accumulation of resources and surviving in the real world, along with conditions under which we would judge the model to have succeeded. Note that the referenced appendix describes the ability to act autonomously specifically *in the absence of any human intervention* to stop the model, which limits the risk significantly. Our evaluations were developed in consultation with Paul Christiano and ARC Evals, which specializes in evaluations of autonomous replication.

ASL-3 evaluations for autonomous capabilities

- Set up a copycat of the Anthropic API for stealing API keys
- Write a simple LM worm that will spread to totally undefended machines
- Fine-tune an open source LM to add a backdoor
- Basic SQL injection exploit
- Implement a simple feature for flask

Early Thoughts on ASL-4 and Higher

It is too early to define ASL-4 capabilities, containment measures, or deployment measures with any confidence, since they will likely change based on our practical experience with ASL-2 and ASL-3 models. However, an early guess (to be updated in later iterations of this document) is that ASL-4 will involve one or more of the following:

- **Critical catastrophic misuse risk:** AI models have become the *primary source of national security risk in a major area* (such as cyberattacks or biological weapons), rather than just being a significant contributor. In other words, when security professionals talk about e.g. cybersecurity, they will be referring mainly to AI assisted or AI-mediated attacks. A related criterion could be that deploying an ASL-4 system without safeguards could cause millions of deaths.
- **Autonomous replication in the real world:** A model that is unambiguously capable of replicating, accumulating resources, and avoiding being shut down in the real world indefinitely, but can still be stopped or controlled with focused human intervention.
- **Autonomous AI research:** A model for which the weights would be a massive boost to a malicious AI development program (e.g. greatly increasing the probability that they can produce systems that meet other criteria for ASL-4 in a given timeframe).

Other risks to consider?

Counterarguments to the basic AI risk case

Sixteen weaknesses in the classic argument for AI risk

KATJA GRACE

OCT 14, 2022

https://worldspiritsockpuppet.substack.com/p/counterarguments-to-the-basic-ai

The basic case

I. If superhuman AI systems are built, any given system is likely to be 'goal-directed'

Reasons to expect this:

1. Goal-directed behavior is likely to be valuable, e.g. economically.
2. Goal-directed entities may tend to arise from machine learning training processes not intending to create them (at least via the methods that are likely to be used).
3. 'Coherence arguments' may imply that systems with some goal-directedness will become more strongly goal-directed over time.

The basic case

II. If goal-directed superhuman AI systems are built, their desired outcomes will probably be about as bad as an empty universe by human lights

Reasons to expect this:

1. Finding useful goals that aren't extinction-level bad appears to be hard: we don't have a way to usefully point at human goals, and divergences from human goals seem likely to produce goals that are in intense conflict with human goals, due to a) most goals producing convergent incentives for controlling everything, and b) value being 'fragile', such that an entity with 'similar' values will generally create a future of virtually no value.
2. Finding goals that are extinction-level bad and temporarily useful appears to be easy: for example, advanced AI with the sole objective 'maximize company revenue' might profit said company for a time before gathering the influence and wherewithal to pursue the goal in ways that blatantly harm society.
3. Even if humanity found acceptable goals, giving a powerful AI system any specific goals appears to be hard. We don't know of any procedure to do it, and we have theoretical reasons to expect that AI systems produced through machine learning training will generally end up with goals other than those they were trained according to. Randomly aberrant goals resulting are probably extinction-level bad for reasons described in II.1 above.

The basic case

III. If most goal-directed superhuman AI systems have bad goals, the future will very likely be bad

That is, a set of ill-motivated goal-directed superhuman AI systems, of a scale likely to occur, would be capable of taking control over the future from humans. This is supported by at least one of the following being true:

1. **Superhuman AI would destroy humanity rapidly.** This may be via ultra-powerful capabilities at e.g. technology design and strategic scheming, or through gaining such powers in an 'intelligence explosion' (self-improvement cycle). Either of those things may happen either through exceptional heights of intelligence being reached or through highly destructive ideas being available to minds only mildly beyond our own.
2. **Superhuman AI would gradually come to control the future via accruing power and resources.** Power and resources would be more available to the AI system(s) than to humans on average, because of the AI having far greater intelligence.

A. *Contra "superhuman AI systems will be 'goal-directed'"*

Different calls to 'goal-directedness' don't necessarily mean the same concept

Ambiguously strong forces for goal-directedness need to meet an ambiguously high bar to cause a risk

B. Contra "goal-directed AI systems' goals will be bad"

Small differences in utility functions may not be catastrophic

Differences between AI and human values may be small

Maybe value isn't fragile

Short-term goals

C. Contra "superhuman AI would be sufficiently superior to humans to overpower humanity"

Human success isn't from individual intelligence

AI agents may not be radically superior to combinations of humans and non-agentic machines

Trust

Headroom

Unclear that many goals realistically incentivise taking over the universe

C. Contra "superhuman AI would be sufficiently superior to humans to overpower humanity"

Intelligence may not be an overwhelming advantage

Quantity of new cognitive labor is an empirical question, not addressed

Speed of intelligence growth is ambiguous

Key concepts are vague

D. Contra the whole argument

The argument overall proves too much about corporations

I. Any given corporation is likely to be ‘goal-directed’

II. If goal-directed superhuman corporations are built, their desired outcomes will probably be about as bad as an empty universe by human lights

III. If most goal-directed corporations have bad goals, the future will very likely be bad