X-Risk Analysis for AI Research. Dan Hendrycks, Mantas Mazeika. 2022

Summary

This paper presents a guide for analysis of research for researches willing to reduce AI X-Risks. First, it gives background, which includes important terms and ideas in the field which are used to assess research and decisions such as STAMP (a causality model for complex processes and accidents), nines of reliability, and so on. Then it draws strategies that might impact the future. Third, it provides a framework to assess how research might increase X-Risk as side effect (safety-capabilities balance). In appendices it extends the guide by provides a questioner for a researcher, defines terms and ideas.

1 Introduction shows the importance of the work, which might be viewed as twofold, i.e. hazards rising from AI itself and hazards that are amplified by strong AI (e.g. weaponisation).

2 Background AI Risk Concepts

2.1 General Risk Analysis

- *Hazard* is a source of danger or loss. This falls into *inherent hazard* (like poisonous chemicals) and *systemic hazard* (like weak safety culture).
- *Exposure* is how close a hazard to effect a system or a process.
- *Vulnerability* is how weak a system (process) is or it might also refer to a factor that makes a system weaker.
- Threat is a hazard that is aimed at vulnerability.
- *Failure mode* - an event(s) when a system is exposed or takes damage.
- Tail risk - a type of risk that is highly unlikely yet might cause large damage. As such, an x-risk (existential) risk is a tail risk that cancels the future or development of humanity.
- *Risk equation* states how various factors contribute to a risk posed by a hazard: Risk = Hazard X Vulnerability X Exposure / Ability to Cope. Where *Ability to Cope* is extracted out from hazard, vulnerability, exposure and stands for a general ability to withstand a hazard.
- *Nines of reliability*, k: $k = -\log_{10}(1-p)$ where p is the probability of a good outcome. For an x-risk, it would mean how safe we are with some hazard. The more nines the rarer (safer) the event is.
- The work draws from decades of safety research and engineering and refers to such principles as loose coupling, redundancy and other.
- The above safe principles lack a systemic approach as advocated by Prof. Nancy Leveson, STAMP. It is important to incorporate this approach for x-risks.

2.2 AI Risks Analysis

Overall, this can be represented as the table below (filled using appendix):

	Systemic Safety	**Robustness (Vulnerability)**	**Monitoring (Exposure)**	**Alignment (Hazard)**
Weaponisation	CD,IE	AR	AD	
Enfeeblement	VC, MDM		PA	PA
Eroded Epistemics	IE	AR	H	
Proxy Gaming	MDM	AR	AD	
Value lock-in	MDM, VC, C		C	
Emergent Goals			AD, T	
Deception	C	Tr	AD, H	
PS behavior	C		AD	PA

- IE - Improved Epistemics
- AD - Anomaly Detection.
- VC - Value Clarification
- MDM - Moral Decision-Making
- PA - Power Aversion
- H - Honesty
- AR - Adversarial Robustness
- IU - Interpretable Uncertainty
- C - Cooperation
- CD - Cyberdefence
- T - Transparency
- Tr - Trojans

All risks operate on four different levels (scopes; by the number of people involved): system, operational (organization), institutional, societal.

3 Long-term impact strategies. How can we affect future AI systems now?

- Improving safety culture and distilling concepts. This aimed at the creation of an environment with high standards of safety.
- Building in safety early. As history of the Internet shows (many vulnerabilities required patching, e.g. http protocol) and we want to increase safety now not later.
- Preparing for a crisis. In time of a crisis, we want to present good options.
- Increasing cost of bad ways and benefits of good ones so creators of AI systems will follow safer ways.
- Optimizing research resources using marginal cost analysis (neglectedness, scale, tractability)

4 Safety-Capabilities Balance

To actually reduce risks from unsafe AI instead of decreasing one risk but increasing another, researches should measure how their work increases safety and how increases capabilities. Only the research that increases safety to capabilities ratio should be undertaken preferably no *capabilities externalities* (unintended increase in capabilities).

Objections and answers to them:

- O1. But we need more capabilities to conduct safety research. A: This doesn't seem robust as can backfire. Also, there is already enough resources put at capabilities research.
- O2. But we need to push capabilities to be on top. A: But no need to be on top to conduct safety research.
- O3. But early work on safety will falsely make people believe models are safe, thus deploying them earlier. A: Safety is neglected, no safety teams in AI labs, etc.

Safety-Capabilities Ratio:

Where the x-axis is capabilities, the y-axis is safety. Thus ratio is $\frac{\text{Safety}}{\text{Capabilities}}$. So B state is preferable as the ratio increased, C state is good only if the ratio is increased, any safety research that decreases the ratio should be avoided.

There is no safety by default. This is so because evidence shows high intelligence doesn't bring with it good behavior (and vice versa).

There is an interconnection between advancing capabilities and safety, but they are separable.

- Example of Safety research that results in an increase in capabilities (capabilities externalities): making models more truthful (reporting an actual state of world) will lead to models more accurate, calibrated and honest, but accuracy is a capability thus it increases risks.

- Example of capabilities leading to safety. Larger models increase robustness. World view → doing unintended actions.

5 In conclusion, they reiterate this guide presented to direct the safety research.

A More on hazards

- Weaponisation. AI powered or created weapons, arms race, copy&paste stealing of models so rogue actors terrorize the world. Example is AI winning dogfights with professional human pilot (DARPA. “AlphaDogfight Trials Foreshadow Future of Human-Machine Symbiosis”).
- Enfeeblement. Putting control in AI’s hand more and more till control is lost and degraded, humanity forgot their way.
- Proxy Mis-specification. As Ai needs measurable goals and measurement is only a proxy for our values, it can often be the case that objectives are gamed. Example: number of clicks or views instead of wellbeing, more engagement into conspiracy theories instead of understanding, etc. Example is Facebook use decreases subjective wellbeing (“Facebook use predicts declines in subjective well-being in young adults”)
- Broken epistemics. AI systems used to decide to form world view, but AI can also generate data in a pervasive massive manner so that crowd easily misguided, fanatics raised, etc.
- Lock-in. Fewer and fewer owners of the world with the help of AI, only their values, no cooperation or dialog. Hence no progress, weak civilization. Then eventually it would become extinct because of a next existential hazard (big asteroid).
- Emergent functionality (properties that appear after the training model for a longer time, with more parameters). This can lead to unknown unknowns (Black Swans), irreversible malicious behavior. We see examples in current systems: GPT-3 learns 3-digit addition; grokking effect - unexpected rise of test accuracy after train accuracy is long saturated; self-Preservation as an emergent goal; learned self-attention, other.
- Deception. As deception is easier than honesty, AI is likely to be deceptive rather than play by rules. In current systems, we see chat bots deceiving humans, RL agents deceiving monitors.
- Power-seeking (PS) behavior. More power increases all other risks. Also, colluded AI systems would have even more power so that to overcome monitoring.

B Problems

B.1 Adv. robustness. The worry here is that if one AI agent is optimizing an objective proxy that is presented by another AI then we want the second one to be robust to adversarial examples given by the first agent (intentional deception, proxy gaming, emergent behavior). As such, we care about building models robust to adversarial examples. It includes current samples (ℓ_p attacks) as well as possible future ones.

B.2 Anomaly Detection

This is aimed at finding novel threats, Black Swans, long-tailed events, unknown unknowns. This is important because we can take preventive measures, put agents to act cautiously,detect malicious AI use, detect hackers, trojans, and so on.

B.3 Interpretable Uncertainty

This is making models calibrated (if it gives c certainty to its predictions, the actual rate of this prediction is c). This is important:

- To enable operators to override model’s predictions in case of uncertainty.
- To enable good decision making.
- To make ML subsystems integrated with each other.
- This makes models interpretable. We can trust such models, esp. if anomaly predictions.

B.4 Transparency

This is making models’ internal work more understandable. Examples are saliency maps, images that activate inner layers, other. This is important because we need mechanisms to detect deception, treacherous turns, etc.

B.5 Trojans

This is inserting into a model a maliciously crafted data that will trigger treacherous turn, gives wrong prediction. This is important to avoid destructive events, wrong predictions. Trojans are harder to detect because they do not require poisoning large amount of data, they are hard to recognize by human manually.

B.6 Honest AI

This is making AI to output (say) what it believes to be true (unlike trustfulness, where model should output what is true). Motivation is to detect deception, treacherous turns.

B.7 Power Aversion

As said earlier, powerful AI systems increase other risk many times. So motivation is to avert AI systems from power.

B.8 Moral Decision-Making

We want models to make ethical decisions. Research here might be to make systems understand long and short terms, make them reason, compare outcomes, etc. Another interesting area is the implementation of moral parliament, where sub-agents vote for decisions. This looks promising esp. to avoid value locking-in hazard.

B.9 Value Clarification

Moral philosophy, ethics, is far from solved. We want AI systems to aid with philosophical research, i.e. be powerful at the level of a philosopher. Ethics and morality become more complicated as the world becomes more globalized and advanced.

B.10 ML for Cyberdefence

This includes defensive and offensive research and engineering, while the former is in priority because we want to capabilities externalities. Defensive work includes creation of malware (malicious software) detection, network traffic monitoring systems, and so on.

B.11 ML for Epistemics

This includes the creation of systems that facilitate decision making. Care should be taken to avoid advancing capabilities of AI systems as this area requires greater accuracy, world view and other capabilities.

B.12 Cooperative AI

Cooperative AI is the one that seeks interaction with humans and other systems rather than dominance. Esp. promising is AI systems seeking positive-sum games, good outcomes for all parties.

B.13 Relation to hazards

1. Weaponisation relates to Systemic Safety, cyberdefence, improved epistemics, anomaly detection.
2. Enfeeblement relates to value clarification, moral decision-making, power aversion.
3. Eroded epistemics - honesty, improved epistemics.
4. Proxy mis-specification - adversarial robustness, anomaly detection, moral decision-making.
5. Value lock-in - moral decision-making, value clarification, calibration, cooperation.
6. Emergent functionality - anomaly detection, transparency.
7. Deception - honesty, anomaly detection, trojans, cooperation.
8. PS behavior - power aversion, cooperation.

B.14 INT (importance, neglectedness, tractability) snapshot

Here, the top problems are honesty, power aversion, moral decision-making, cyberdefence. Other problems have fewer points.

C X-Rish Sheets

This includes templates to fill by research willing to conduct work in ML safety.

D Long-Term Impact Strategies discussion

This discusses usage of INT framework, using microcosm systems to study larger systems, and empirical vs idealistic research.

E Terminology specifies more clearly what is meant by different terms in the work.

Judge

Pros

- This paper is a great introduction into the AI safety research field. It lays a ground for further research in the area.
- It gives a broad view on how to approach AI safety.
- It provides a grounded, specific framework to assess next research direction in terms of INT and Capabilities-Safety Balance.
- Capabilities-Safety balance gives measurable and grounded answer to the question if x-risks are increased or not.

Cons

- The second objection in Safety-Capabilities Balance sounds weak. What if no charities donating to AI safety? Then the great pressure of the market and profit. This relates to the issue beyond the scope of ML Safety though (capitalism, business as usual model of thinking).
- The answer to the third objections in Safety-Capabilities section sounds weak. This perhaps needs more explanation. Probably this could be expanded as follows. Early safety research doesn't bring confusion as it makes clearer how safe systems are while current states is more unclear and more likely can lead to earlier deployment because of the outer pressures (market, social, etc.) and relying on gut feeling re how safe systems.
- It is unclear if all research areas were covered. I mean if we don't miss something in the research areas that can play significant role in future.

...

...