Something Something Meta Review of AI Failures and AI Failure Categorization

Max Williams
University of Louisville

Roman Yampolskiy
University of Louisville

2020-09-24

Abstract

Oh gee what to even say. AI failures are becoming like totally a legit thing now. This field has like the totally out there wako people then there's also the basically engineering side of things. Just scraping though news you can find some pretty worrysome stuff. That's half the reason we're here, to show these failures before these systems get too out of control.Yeah. So I'm going to discuss how we have and should organize AI failures, then combine all the lists that I can find to make my own mega list.

1 first section

By combining our existing knowledge on bad things that have happened and our wild guesses about things that could happen, we can really do some solid guessing toward what kind of problems we could see in the near future. Far future, maybe not, and this really doesn't work for AGI because it'll probably get completely, categorically different from current AI. Like currently we're like uh oh we put computers in charge of stocks and they blew it up. The point being that WE put them in charge of stocks, and they stayed in the box. Once the AI can leave the box we assign it to (ever for in-box related goals) then these kind of failures become the only-make-them-once sort of failures.

2 existing work

3 paper notes

[6]

20 pages, focused on rl training safe driverless card. Problem is that evaluating rl agaents can take longer than training them, so add an adversary that evaluates them on the hard stuff, like a driving instructor skipping straight to tight parallel parking when they see you're good

Doesn't actually list any failures

[7]

33 pages. I like this already, taking the analogy of organizations as agents seriously.

This is a really complex paper, beyond me in a lot of ways but really neat.

[1]

20 pages. This is basically one big failure example, about a specific technique called "adaptave control". It also includes details about an overreaction to these failures, followed by "don't use it unless it's applicable" advice which has to be hammered into the naive buyers of COTS AI software.

[3]

5 pages. "Post-accident attribution accident to a 'root cause' is fundamentally wrong."

Outch, okay yeah I get it. Damn.

[2] [5] [4]

TODO get yam's papers in here

3.1 categorizations and their data

4 unified categorization and how much data we got yeeeah

So I like the tag based system, and title + link + short description is a pretty good way to go. The tags all seem pretty distinct and nessesary, but are tricky to convert between.

References

[1] Brian D. O. Anderson. Failures of adaptive control theory and their resolution. *Commun. Inf. Syst.*, 5(1):1–20, 2005.

[2] Stephanie Carvin. Normal autonomous accidents: What happens when killer robots fail? *Carleton University*, 2017.

[3] Richard I. Cook. How complex systems fail. *Cognitive Technologies Labratory*, 1998.

[4] Nargiz Humbatova, Gunel Jahangirova, Gabriele Bavota, Vincenzo Riccio, Andrea Stocco, and Paolo Tonella. Taxonomy of real faults in deep learning systems, 2019.

[5] Matthijs Maas. Regulating for 'normal ai accidents': Operational lessons for the responsible governance of artificial intelligence deployment. pages 223–228, 12 2018.

[6] Jonathan Uesato, Ananya Kumar, Csaba Szepesvári, Tom Erez, Avraham Ruderman, Keith Anderson, Krishnamurthy Dvijotham, Nicolas Heess, and Pushmeet Kohli. Rigorous agent evaluation: An adversarial approach to uncover catastrophic failures. *CoRR*, abs/1812.01647, 2018.

[7] Rodrick Wallace. Failure of real-time multi-component, multi-level cognitive systems on clausewitz landscapes. *The New York State Psychiatric Institute*, 2018.