Case Studies on AI Safety Issues

by Darshil Dhameliya

In the 21st century, modern computer science and engineering advancements have led to one of the penultimate use-cases of all the sciences: potentially creating an artificial intelligence vastly smarter than humans. Such an advanced AI System is sometimes referred as 'Artificial General Intelligence,' or AGI in short. But these recent breakthroughs in the field of AI have also provided a glimpse to the emergent hazardous behavior, something these complex systems are not expected to act like. Let's look at some examples.

Issue One: When It Comes to Gorillas, Google Photos Remains Blind [1]

This article reveals a shocking peek into how 'the future' has turned out to be. A few decades ago, researchers working on the field of computational intelligence wouldn't have realized we would come up with models that can aid in all sorts of analyses: processing texts, images, audio recordings, videos and all kinds of data. It was an alarming signal when some people saw the Google Photos Engine classifying Black people as "gorillas."

The response by Google (now a subsidiary of Alphabet Inc.) was that they were working on a long-term fix. But 2 years later, Google just ended up removing some labels like 'gorilla', 'monkey', 'chimpanzee' from the model completely. Google researchers just couldn't figure out where the model was going wrong in falsely classifying such images. Google Photos now serves more than a billion users and has trillions of images available in its database. [2] [3]

Even though the users of the service take photos in varying settings, and with a lot of training data available, the model still gravely misclassifies in some outlier case. It is fair to assume that frequency of such misbehavior of AI models are only going to increase as we build more complex models which are even harder to interpret. Some researchers worry these Deep Learning based models boost bias related to gender or races, for instance. [4]

While a simple task like classification doesn't have physical implications, think about some of the high-stakes automations like self-driving cars being developed by Waymo (also a subsidiary of Alphabet Inc.) and Tesla. If such a model wrongly detects a living being as a non-living object and does not stop itself from hurting someone, that can have serious moral implications. Something like this hypothetical incident also generally gets down to agency and asking who's to blame, but that does not really go to the root of the problems.

As these hazardous behaviors caused by these complex systems only begs the question how to resolve such representations generated by these models. Many researchers in the field are already working ways to tackle such related to fairness. [4] Some approaches involve integrating interpretability tools in the models to better understand how it makes decisions. We can also train the models to detect anomalies by potentially using low density estimates created by these predictive models while processing sparse data points or the outlier observations.

An image can be modified in a way that is not visible to human eye but can make the algorithm wrongly classify it as something else. One such adversarial example is shown in Image 1 where the model classifies cat as something completely different. One approach is to train the model with adversarial examples in classifying them correctly. Such training has shown robustness improvement in models against adversarial attacks. [5]

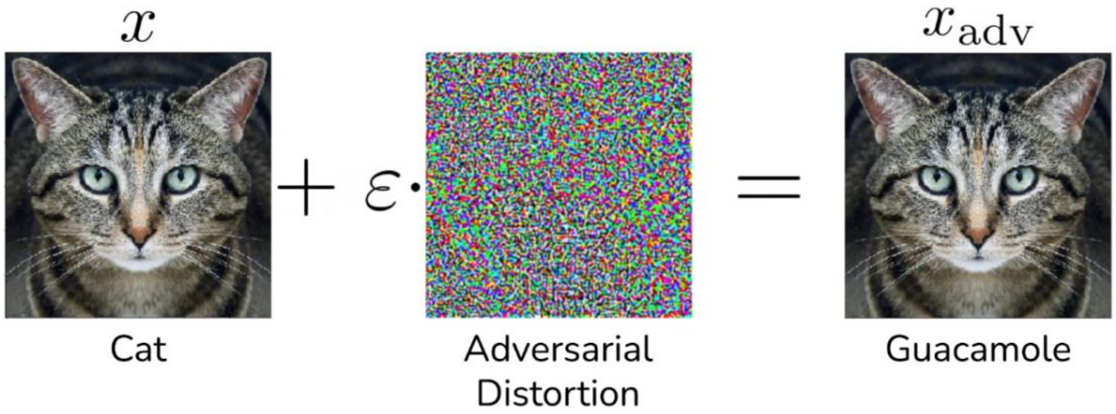

Image 1: An adversarial example where a cat is identified as “guacamole.”

Issue Two: Specification gaming: the flip side of AI ingenuity [6]

Let’s consider a different class of problems where Reinforcement Learning (RL) agents behave in unintentional ways. These type of misalignment in objective specification is analogous to Midas’ Touch where everything Midas touches turns to gold, including food!

Imagine an environment like in Image 2, where the robotic arm is supposed to put the red Lego block on top of the blue one. But the task specification is something like ‘bottom face height more than that of the red block.’ So, the agent shows its ingenuity and just flips the red block on its top! This type of eccentric behavior of artificial agents is also known as ‘Specification Gaming.’ If the task also specifies keeping top face high, that works as desired.

Specification gaming can also lead to positive feedback or creative outcomes. This happens when the environment or the simulation in which the agent learns, is well-defined. Consider DeepMind’s AlphaGo exploiting a well-defined game with clear constraints, the Chinese game of Go. In this setting, AlphaGo has led to some exceptional moves surprising even the best players of the game as they had never observed or thought of such moves. [7]

This doesn’t mean that a well specified environment is enough for positive outcomes in specification gaming. The task/objective specification also need to be aligned with the intended outcome. RL agents tend to exploit the rules to achieve exactly what they are specified to do but if the intent isn’t properly integrated in the reward/feedback design, the agent can lead to disastrous outcomes. There are a lot of examples for such behaviors. [8]

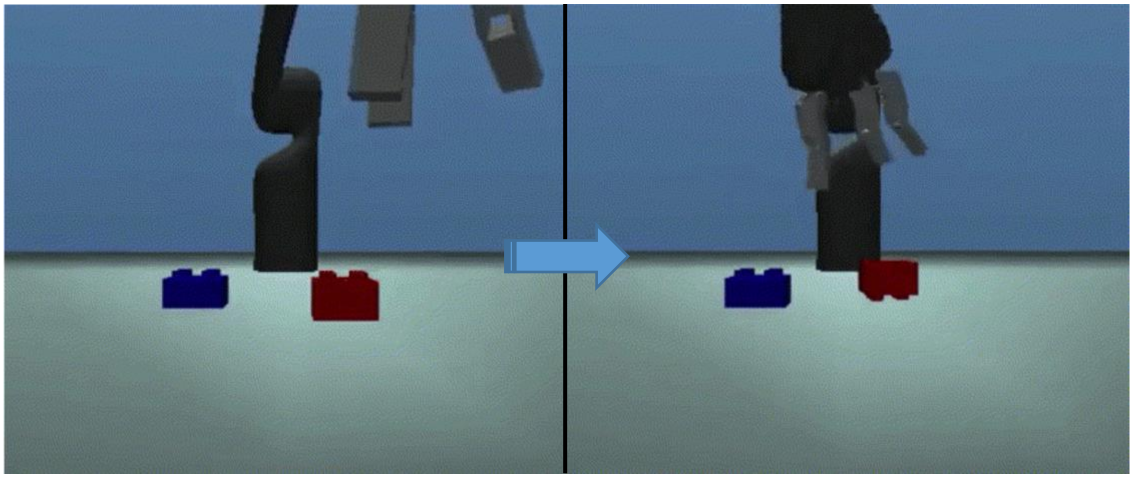

Image 2 Source: Popov et. al., Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation (2017)

Most of the times, some model/simulation of the real environment isn't well-defined because it's just not easy for the designers to get every detail right. These simulation bugs can also lead to poor translation of optimal actions when the agent is deployed in a real-world scenario. The agents tend to behave in new ways maybe because they don't follow usual human rules, intuitions or biases existing in our brains due to evolutionary processes.

Some proposed solutions in response to such misspecification of reward functions is to let the agent learn human preferences through human feedback. OpenAI is experimenting with such reward shaping so that the agents might learn desired optimal policies and better aligned intent specification. [9]

There are many philosophical inspections that need to be assessed to find potential loopholes in more advanced systems. Look at ELK by Paul Christiano et. al., for example. [10] Since we haven't seen such models behave in front of our eyes yet, many philosophers have come up with previously unexplored thought experiments, e.g., a sufficiently powerful agent could temper the user preferences, subtly changing the goal itself to achieve high performance measurements.

Aside from the philosophical and technical considerations, we need corporate cultures to focus more on AI Safety related concerns. Such research is not really incentivized by the modern free-market dynamics. In fact, current incentives create friction against moving towards a positive direction. So, a lot of regulatory input is required from the authorities to fix this misalignment of monetary incentives leading to existential risks related to AI.

References:

[1] https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[2] https://blog.google/products/photos/storage-changes/

[3] https://www.theverge.com/2019/7/24/20708328/google-photos-users-gallery-go-1-billion

[4] https://arxiv.org/abs/1607.06520

[5] https://arxiv.org/abs/1902.06705

[6] https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

[7] https://www.wired.com/2016/03/googles-ai-wins-pivotal-game-two-match-go-grandmaster/

[8] https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml

[9] https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/

[10] https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit