Trustworthy AI Assessment List
Feedback from UC Berkeley Center for Human-Compatible AI
11/29/2019

We think that the Trustworthy AI Assessment List is an excellent starting point for thinking about how to increase the trustworthiness of AI systems. We'd like to propose some improvements with reference to a domain in which we expect significant issues to arise: algorithmic trading. As AI technology advances, there will be strong incentives to develop AI systems which not only implement pre-specified trading strategies, but also learn novel strategies of their own.

However, some profitable strategies qualify as illegal *market manipulation*; and some may lead to *market instability*, as already seen in various flash crashes. The legal definition of market manipulation is complex, and by default will be difficult to communicate to AI systems. Since finance is such a data-rich field, and there are such strong monetary incentives to develop novel trading strategies, we expect trading AIs to showcase examples of untrustworthy behavior which will presage issues in other industries. This possibility highlights several ways to improve the assessment list:

Technical robustness and safety

The concern that an algorithmic trading AI may learn to engage in market manipulation does not naturally fall into any of the headings in the Technical robustness and safety category. Market manipulation would not necessarily be caused by adversarial attacks. Nor would it arise in response to unusual or unexpected situations, since it would be the AI system itself causing the unexpected situation to occur. Nor is this a concern about dual-use technology, since this behavior might arise without human operators noticing or endorsing it. Rather, the problem arises from the system exploring new options which are superficially appropriate, e.g., legal trades meant to capture value for the agent's owner, but cumulatively harmful or illegal, e.g., indirectly causing a stock's price to increase followed immediately by selling it.

We propose adding a new subcategory, *Alignment of system incentives*, including the following questions:

- Did you precisely specify what qualifies as desirable and undesirable behavior from the system?
- Did you consider potential negative consequences from your AI system learning novel or unusual methods to score well on its objective function?
- Did you consider how the incentives of your AI system might interact with those of other systems deployed in the same environment?

Human oversight

The example of AI traders also highlights some of the difficulties involved in human oversight. Undesirable outcomes may occur very quickly, or else involve such large-scale patterns in the data that humans cannot immediately identify problems. Another challenge to human supervision,

especially in algorithmic trading, may be computational intractability -- there may be just too many separate transactions for a human to oversee them all, even when the undesirable outcomes are "small-scale" enough to identify from individual transactions.
Even when problems have been identified, a "stop button" may be insufficient: for example, immediately ceasing trading may exacerbate market distortions.
We propose adding the following questions:

- Have the human overseers been trained to monitor the AI system and interpret its behavior? Have you tested whether they can reliably recognize undesirable behavior?
- Did you consider the potential consequences if human oversight were absent for short periods?
- As an addition to "Who is the "human in control" and what are the moments or tools for human intervention?": Can those tools monitor AI behavior which occurs at timescales too fast or in excessive quantity for humans to process?
- Instead of "Does this procedure abort the process entirely, in part, or delegate control to a human?": "Does this procedure abort the process entirely, in part, or delegate control to a human or default safe AI system?"
- What assumptions about humans/the environment/other systems deployed in the same environment does the AI system use when making decisions? Can these assumptions be made explicit and be manually checked by an (ideally human) overseer?

Accountability

If market instability (such as a flash crash) occurred due to AI misbehavior, the losses involved might be significantly beyond the ability of most companies to redress. With that in mind, we propose adding two additional questions to the Accountability section:

- If harms might potentially go beyond your own ability to redress, have you explored ways of increasing your ability to do so, such as purchasing insurance?
- Did you ensure that the system's ability to affect third parties is as limited as possible, and that it only has access to actions which are necessary for its intended functionality?

While in this submission we have focused on the example of algorithmic trading AI, we expect that similar challenges will arise for many other applications of AI. The questions we have suggested should provide useful guidance for deploying AI systems in a range of domains, such as conversational AIs and recommendation systems, and particularly those which make use of reinforcement learning.