AI Race Safety through Agile Risk Management

Bogdan Burlacu, John Diamant, Dave Graham, Steve Kommrusch, Colby Renfro, Scot Weber, and Dan Yee

The dynamic nature of AI advances will require a risk management system that can be as agile as AI technology advancements. We propose augmenting the work of governments and standards committees with a market-based approach to mitigate AI race safety risks. The core concepts of our proposal are: establishing liability laws governing AI, requiring AI malpractice insurance, and incentivizing independent 3rd party risk assessment. Our proposed agile system will avoid unnecessary limitations to rapid AI development and also react appropriately to new risks.

Groups such as OpenAI, GoodAI, the Future of Humanity Institute, and others have contributed to the field of AI safety. As an example of the visibility this is receiving, the International Organization for Standardization (ISO) is working to define standards for AI with ISO/IEC JTC1 subcommittee 42 focussing on artificial intelligence [1]. While standards are helpful, a system that results in safety companies like Underwriter's Labs based in the USA or TUV based in Germany may allow the marketplace itself to react to new ideas and approaches quickly. To encourage risk to be appropriately considered in product development, we propose these steps.

1. Clear liability rules including appropriate 'required standards of conduct' need to be put in place for AI products. Such issues as assigning liability to software subcomponents should be included. Details of these regulations are outside the scope of this paper.
2. Governments should require companies producing products which include AI to have AI malpractice insurance. Underwriters will set the minimum price based on fines and penalties assessed by courts or government agencies along with the assessed risk of the products being insured.
3. Governments and courts should weigh 3rd party evaluations of products in decisions and penalty assessments regarding AI safety. These judgments should be clear enough to encourage a market response. For example, a finding of negligence in the design of a self-driving vehicle should be less common when serious issues found in a 3rd party audit were addressed, and the penalty should be less severe.

Given the requirement for malpractice insurance and the legal attention given to 3rd party evaluations, either the insurance companies or testing companies would be relied upon to evaluate product safety in alignment with government regulations but also with perceived future risk for a given new product. Such evaluations may include ensuring robust code testing techniques, quality training data sets, testing for distributional robustness on inputs, safe environment exploration when a robot is unsure of the best next step to a goal, and other new techniques for safety as the community develops them.

Without the requirement for AI malpractice insurance, the financial (and societal) risk taken by companies is unlikely to be accounted for. In the race to develop a product, companies are likely to take risks that would result in costs too high to bear if a serious negative consequence occurs, resulting in bankruptcy for the company and damage to society. By requiring insurance, the risk is born by the insurers who will charge based on their assessment of the safety of the product (including the testing done on it), which will encourage the creation of a market-based watchdog system. This system, guided by government regulations and legal precedents, should be more responsive to changing techniques in the AI field than solely relying on standards committees. Additionally, the 3rd party evaluators can improve their testing strategies based on knowledge gained when testing prior products while preserving confidentiality for their clients. This confidentiality is a key benefit of our proposal as some governments have rules limiting their ability to protect commercial confidential information. Another benefit of 3rd party evaluations is that they could be leveraged to a consumer-visible AI safety market signal [2] once their reliability is established.

Established product liability covers some of the existing products that are or will be augmented with AI technology in the near future. However, we recommend liability be extended to include software in such a way that End User Licence Agreements cannot override protections against negligence. In the EU, GDPR (General Data Protection Regulation) has up to 4% annual revenue penalties for violations, but the penalty is reduced to 2% if you can show you followed GDPR practices [3]. Such penalty reduction is what we are proposing in point 3 of our proposal. In the US, the rules for a company being found negligent include: "...failing to conform to the required standard of conduct..."[4]. We expect through our proposal that "required standard of conduct" will come to include 3rd party evaluations of product proposals. Given the incentive of the insurance companies to properly assess risk, the evaluations would include not just likelihood of a poor design behaving incorrectly, but the potential damage that could result if a product performed incorrectly.

Our proposal to manage the consumer product field during development can have broad effects. Industry safety requirements will help guide non-profit research, and even military uses of AI will benefit from safer AI practices in industry as recently the military is relying more heavily on innovation in the consumer sphere than it has in the past [5]. This proposal additionally creates an AI safety acculturation which will infuse good practices into most members of the AI community, even members not directly involved with the insurance system proposed.

There are enormous challenges and opportunities for advancing artificial intelligence. Our proposal aims to build upon AI safety regulations being considered today by governments and standards organizations. The goal of our proposal is an agile market-based system which can react quickly to changing AI techniques as advances in the field occur. By establishing AI liability laws, requiring AI malpractice insurance, and incentivizing 3rd party risk assessment, AI development can proceed efficiently but with appropriate safeguards against producing harm.

References

[1] Artificial Intelligence: ISO/IEC JTC 1, Subcommittee 42
https://www.prnewswire.com/news-releases/national-experts-sought-for-new-standards-committee-on-artificial-intelligence-isoiec-jtc-1-subcommittee-42-300583517.html

[2] Market signalling for economic information exchange
https://en.m.wikipedia.org/wiki/Signalling_(economics)

[3] GDPR penalty assessment rules
https://www.imperva.com/blog/2017/03/gdpr-series-part-4-penalties-non-compliance/

[4] US negligence law:
https://legal-dictionary.thefreedictionary.com/Negligence+(law)

[5] Naval research relies more on consumer products than in the past
http://blogs.berkeley.edu/2017/10/11/office-of-naval-research-goes-lean/