Tackling the perils of dual use in AI

Considering the potential for unintended harmful applications of AI tools can lead to deeply concerning findings. An urgent question is how to achieve the right balance between keeping science open and preventing misuse or malicious repurposing.

When chemists at Stanford University submitted a research proposal that involved the use of AI to predict the toxicity of chemicals and materials, questions raised by the university's Ethics and Society Review panel gave them pause for thought, as Shankar and Zare write in a Correspondence in this issue. Naturally, the researchers wrote the proposal with beneficial purposes in mind, namely for developing materials that are safe to use. But they realized that the same AI approach could be used for harmful applications, involving the development of toxic materials.

This story is not unlike the one from drug discovery researchers who recently wondered what what would happen if a generative AI approach to find molecules that are beneficial to health were given the opposite goal, namely to find molecules with high toxicity. The disturbing answer, as reported in a recent Comment by Fabio Urbina et al., was that, within 6 hours of computing time, the algorithm found 40,000 candidate toxic molecules, including the nerve agent VX, but also many new and even more toxic compounds together with their precursors. The realization dawned that as AI can be harnessed for a particular beneficial effect, the goal may be flipped and the algorithm could just as easily be exploited for a harmful effect.

Of course, the findings are only computational results, and the next steps for anyone with harmful intentions, who wants to synthesize, store and transport the new chemicals, are not trivial. Still, the work indicates the worrying possibility of new routes to circumvent chemical controls and watch lists.

The authors themselves were sufficiently shocked that they quickly shut down their computational experiment and kept the results, and knowledge of dangerous compounds and their precursors, under lock and key. However, the Comment attracted considerable media attention and as a result the authors found themselves having many stimulating conversations, with interested scientists, security experts, journalists and others, and they came to realize it would be worth going back to their results. "There could be an opportunity to do more analysis, learn more and do something positive", corresponding author Sean Ekins told us. This could involve finding antidotes for nerve agents and discovering useful information to help detect illicit chemicals.

The authors are now more optimistic that the research they embarked on could be turned into something 'good'. At the same time, although further research and collaborations seem promising, it is clear that steps are necessary to avoid giving easy access to the detailed methods. But replicability, reproducibility and reusability are vital for progress in machine learning and for its application to different scientific fields. The question of how to responsibly share code and data in this scenario will be an important one to tackle. For most machine learning research, the advantages of providing open, unrestricted access to the data will outweigh any possible distant risks. But when positive applications can be easily inverted and repurposed for negative ones, restrictions on access to data and models seem to be required.

Not publicly sharing data or models at all would certainly avoid any risks. But this approach also makes the validation of the original results extremely challenging. Moreover, as materials will then have to be shared upon request, a significant burden is placed on authors, who will need to decide on a case-by-case basis whether access should be granted.

Alternatively, this responsibility can be shifted to institutional organizations, which can restrict access to data and models, while allowing researchers to submit a request for access. Video data collected in the Databrary project, for example, is collected by several institutions, and they can each authorize their own researchers to access other parts of the database. Ethical oversight is ensured by the Institutional Review Boards and ethical use of the data.

A common practice for some fields is to offer access to their trained method via a web server or API, such as the REDIAL project to find anti-COVID drugs. This can ensure that a method is only used as intended and that retraining is prevented, but it also restricts further development unless the code and original training data for the method are available in addition, as is the case for REDIAL.

OpenAI famously restricted access to their GPT-3 API via a waiting list. Although the model can now be used directly after the creation of an account, the automated use through their API is subject to a strict user agreement and prohibitions — such as using the output of models directly as social media bots. OpenAI also offer their own filter system to detect toxic output of the model and alerts users that, should they choose to still view the questionable output, they should not share it.

Another possible route, considered by Fabio Urbina et al. for future work, is federated learning, in which data can be kept secure and model development is distributed. An example for the use of federated learning in bioinformatics can be found here. If the participating institutions agree to keep their federated learning pipeline open for other groups to deploy their own models, then the original methods can be replicated or improved without disclosing any of the training data. However, this still requires that the aggregated models are either harmless or secured.

As Shankar and Zare conclude in their Correspondence, there is an inherent and non-trivial conflict between making AI research public and protecting it from abuse and misuse: we join the authors' call for action, which invites researchers, policy-makers and interested parties to come together, ideally at dedicated conferences, to devise a workable plan to overcome this dual-use problem. ❐

Published online: 22 April 2022
https://doi.org/10.1038/s42256-022-00484-6