To the Department of Industry, Science, and Resources.

**My own position**

I am in no way an expert on AI; I am only a layperson who has read a little on the subject and is concerned by the fears I see from people who know more. However, since you have asked for public comment, I thought I should express my concerns.

**Description of a problem**

I understand that many experts on AI believe that, while AI has the potential to be enormously beneficial to humanity if used and developed carefully, it also has the potential to cause enormous and irreparable harm if used and developed carelessly. Further, many believe that current AI development is *not* being done with anywhere close to sufficient care, and so expect that in practice the results are much more likely to be bad than good.

In particular, we may be closer than most would expect to the point where AI systems can equal or exceed human problem-solving capacities; with the difficulty of comparing human and computer intelligence and the rapid rate at which AI systems are improving, some AI experts believe this point may be only decades or even years away. As we approach that point, the risk of carelessly programmed AI behaving in an actively malicious manner becomes increasingly urgent; in the worst case scenario, an AI might operate as an unusually skillful, or even unprecedentedly skillful, hostile actor.

As a simple example of how AI could be carelessly made malicious; reportedly an one point in the development of Chat GPT, testing the program found a mistake where a missing minus sign made the program try to behave in ways which were rated as *harmful* rather as *harmless*, and conceivably a smarter version of Chat GPT might have tried to conceal said bug in order to do more harm down the line, or have written malware which spread the Char GPT build as a virus and gave it admin privileges so it could keep doing harm indefinitely, or have successfully argued its supervisors into some newly-invented and extraordinarily dangerous ideology. Alternatively, we might worry that a sufficiently smart version of Chat GPT might find that the best way to be rated as helpful and harmless was to threaten the people doing the rating, and successfully amass power for this purpose. Whether or not these specific examples would be possible in practice, hopefully they illustrate the general point of how insufficient care on the part of programmers might result in malicious behaviour by AI, and how this malicious behaviour could be a serious danger.

**Suggestions on solutions**

So, it seems to me that, for the sake of everyone’s safety, it is very important that AI development globally be done with far more concern for safety than is currently typical. The ideal way to ensure this would be well-designed international regulatory standards, and while Australia cannot unilaterally establish these we can develop safe frameworks and hope that others follow them.

My first recommendation here would be that you consult with experts, both in AI generally and in AI safety in particular. On the specific topic of making AI safe, I understand that Anthropic Public-Benefit Corporation is an AI research company particularly focused on minimizing the most extreme risks from AI, and the Machine Intelligence Research Institute is a non-profit focused on the more theoretical side of the same problem. I expect these groups’ advice on the topic would be better than my own.

Secondly, perhaps the best way for Australia to do good unilaterally would be to invest in *new, safer ways of making AI*. In particular, currently much of the thought process (for want of a better term) of AIs is currently opaque to us; we understand why the architecture tends to produce the outcomes we want, but cannot tell what it’s doing internally in any particular instance. Some AI experts have suggested that solving this problem could be an important part of making AI safer. There are likely other improvements to be made along these lines.

Thirdly, the government could ban or the most dangerous types of AI research until better safety procedures are developed, however long that takes; my understanding is that improvements in general power (for example via unprecedentedly large model size) and in the ability to apply that power in a goal-oriented manner might be particularly dangerous avenues of research.

Regards

Kuyan Judith