The Difficulty with Regulating Generative AI

Much of the hype surrounding AI at present can be attributed to the recent advances in generative AI technologies (ChatGPT, DALL-E, etc). These technologies are impressive in their way and are not without their uses, but their reliance on Deep Learning (DL) places a hard limit on what they can accomplish and renders them difficult, if not impossible, to effectively police. The essence of the problem is that AI created using DL methods has no potential to ever understand what it is producing and due to the opacity of the model, human operators cannot make sense of where it went wrong to correct it in any way.

One of the main issues to address is that of the generation of inappropriate content: for example, a chatbot explaining to inquiring individuals how to build a bomb, or an image generator creating deep fake imagery of a public figure behaving inappropriately. If the machine had anything approximating real understanding, it could assess the appropriateness of the input itself. Or if the model wasn't an opaque mess of a trillion nodes, a developer could design it not to answer questions like this. But neither of these are within the realm of possibility for any model developed using DL-methods, so how can the content produced by such an AI be expected to stay within the bounds of the regulations we've decided upon?

The current solution is to prepare moderated responses to specific inputs that it has occurred to the developers to flag, which essentially amounts to slapping a band-aid on the problem. As long as this is the only solution available for moderating content creation, motivated users will always find a way around it, whether it be "jailbreaking" the technology

(something chatbot implementations are especially susceptible to) or simply coming up with a way to frame the input without setting off any flags.

One possible solution is to place heavier regulation on the sort of content included in their training data sets so that the resulting model does not possess the capacity to generate inappropriate content regardless of how persistent the user is. The problem with this approach is that it can be quite difficult to anticipate what sort of data could prove potentially destructive in advance. And who would be deciding what is and is not appropriate? Would there be a committee in charge of making these decisions? That begins to have a bit of an Orwellian ring to it. And what about models trained overseas? Would they be subject to the same rules? And suppose we somehow succeed at this task and manage to dictate a perfect training set to create a model incapable of providing content we've deemed inappropriate: at this point we've probably crippled the model to such an extent that entire fields of knowledge have been excluded from it, which severely limits its usage and may result in Australia rolling out inferior AI products compared to the rest of the world.

Though these technologies are responsible for much of the recent hype in AI, their regulation is difficult and they ultimately represent a technological dead end. Overreliance on these technologies in their current form is both irresponsible and societally damaging. Much more dependable technology is needed before we can even think about applying it to anything with real impact.

A different approach is needed if we are ever to overcome the restrictions of the DL approach. We are currently developing Active Structure, an entirely new approach to teaching a

machine language using a self-extensible, undirected network of semantic structures. Through mapping the relations between words, meaning itself begins to emerge. This is one point where our approach distinguishes itself from the prevalent DL-based methods: our approach actually has the potential to understand not only the question it is being asked, but the response it puts together. Our approach focuses on understanding the workings of the unconscious mind and emulating its process of language comprehension in the form of a machine. Our goal is to develop a machine with the comprehension abilities of a human combined with the processing power of a computer. In this, we hope to overcome the failings of currently prevalent machine learning approaches while surpassing the processing limitations of the human mind.