Using ChatGPT safely

even though it's dumb and dangerous

by Dr Simon Thompson, Head of Data Science, GFT UK

This article outlines some harms and benefits of applications using ChatGPT. By identifying a set of harms, we can give a simple methodology for determining if it's ok to use ChatGPT in an application. Some rules for the implementation of the applications that are developed are provided. This article is 100% human-created (apart from the stable diffusion image below, and the screenshots from ChatGPT—obviously)!

Dr Simon Thompson
Head of Data Science
GFT UK

Technologies can break loose from their creator's control. Mobile phones are a great example. Bell Labs invented them, and for thirty years the telecoms industry saw them as extensions of landlines. Their view was that a mobile phone would let you make and receive phone calls when you were away from a fixed line phone. In the 1990s, telecoms discovered mobile messaging (SMS) and a revenue bonanza turned companies like Orange, Verizon, RIM and Vodafone into technology giants. Then the iPhone was introduced and suddenly phones became portals to an internet of applications and services.

Mobile social media bloomed and mutated into something that no one at Bell Labs would have either recognised nor countenanced, and the rest is (ongoing) history. The telecoms industry was out of the picture and Apple and Google now define the product and use cases for mobile phones.

This kind of escape can have unexpected and unwanted consequences. Many have argued that some of the new applications of mobile phones are hugely damaging. For example, applications such as Instagram and Snapchat (consumed via 'phones') are sometimes blamed for destroying young-women's self-image, leading to a rise in suicides[1].

ChatGPT is a new and powerful technology that is also escaping the control of its developers. Until a few months ago, chatbots were narrow and brittle interfaces onto information sources. Applications employing them were expensive to develop and often failed in practice. For example, Facebook launched a service called 'M' in August 2015.

This was slated to act as an automated personal assistant, but in practice, by 2018, Facebook was only able to achieve a 30% automation rate for the service.

In contrast, ChatGPT is a chatbot that has broad abilities in terms of handling different types of conversation and conversing on different questions. It appears to be very flexible and can be adapted to a variety of tasks cheaply and quickly.

Illustration: Mobile phones from the 1950s imagined by Stable Diffusion AI image generator, when mobile phones were just for making calls!

Because ChatGPT has become available and the methods used to build it are so well-known, there is now a rush from other technology providers to launch similar technology, and a legion of independent developers is creating applications using it. We have to get used to the idea that generative conversational assistants now exist and all sorts of people are using them. This paper tries to answer the following questions: So what? What applications are off limits? What should be done to make things safer?

So what?

These tools are underpinned by a generation of machine learning models collectively known as Large Language Models (LLMs). The first of these was the BERT model [2] in 2018 and since then LLMs such as ChatGPT have grown to be 1000 x larger. As LLMs have grown, they have become more and more capable of impressive feats of natural language processing. They are now able to generate long form text, poetry, computer code and interactive conversations.

This is a bit of a shock. There are several reasons why it has happened. Firstly, huge amounts of money have been poured into the effort, allowing a 'brute force attack' with vast amounts of cloud compute used to power LLMs. The training methods developed for the first generations of LLM have also turned out to be very inefficient, and very clever people have created much more efficient ones.

Finally, and possibly most troubling of all, the life blood of machine learning models is data, and the creators of LLMs have (with some honourable exceptions) demonstrated rapacity and ruthlessness to get it.

Despite what it says when you ask it, ChatGPT appeared at the end of November 2022. See illustration. Already, just a few months later, many unexpected applications for ChatGPT are appearing. These range from the apparently ethically unproblematic to the obviously malicious.

In this article, I shall review the application areas that are emerging and outline why these might be considered problematic (or not) and then I will outline what steps can be taken by business people to decide if an application is appropriate and ethical, and what mitigations can be put in place to make fielding something possible.

Illustration: ChatGPT thinks it was released two years before it actually was, apparently because it is powered by GPT3.

We'll cover some relatively benign examples, including:

- Assistive text creation for business purposes
- Text creation for personal purposes
- Coding assistance

After looking at these, we'll present other more problematic examples:

- Students using ChatGPT to cheat
- Disinformation campaigns
- ChatGPT powered apps providing dangerous advice

Having examined these applications and the issues that can arise from them, we list the detrimental ways that ChatGPT may check if an application will ultimately be harmful. Finally, we list some extra safeguards that every ChatGPT powered app should include.

Text creation for business purposes

Some people who struggle with written communication find ChatGPT supportive. In the linked story, a swimming pool installer is able to use ChatGPT to improve his communications with his customers. In this way, he is able to reduce his anxiety about annoying or alienating a customer and feels that he is able to appear professional and competent. My opinion is that this is no more ethically problematic than using a spell checker or wearing glasses to read.

Similarly, using ChatGPT to create long or short form versions of original content for websites or tweets (or even blogs) seems legitimate. But of course, challenges with this approach emerge in how it is used. Getting ChatGPT to write a speech based on a set of ideas that you prompt it with is one thing, getting it to write emails or LinkedIn messages to bombard sales targets is quite another. The ethical challenge here seems to be mass production and the deceptive creation of an apparent empathy and relationship with the target audience. The mass production of personal communication threatens to destroy channels of open communication. How can bots be banned when it is impossible to detect that they are bots?

One huge issue with text creation for business use is copyright.

In some business applications and contexts, concern over copyright is probably overblown. It is very unlikely that the swimming pool contractor mentioned above will face a lawsuit from someone because snippets of a copyrighted document appeared in his or her response to a client enquiry. In other cases, using ChatGPT-generated text could be really problematic. Recently an 'EdTech' company was sued by a publisher for reproducing and redistributing material.

Also, The Financial Times sued a US asset manager for allowing employees to access copyrighted material without payment.

ChatGPT as a content generation engine may create similar cases (Figure 1). Whilst it may be unlikely that a business application based on ChatGPT would create the significant damage and harm that leads to a lawsuit, it is definitely something that businesses must bear in mind when using ChatGPT.

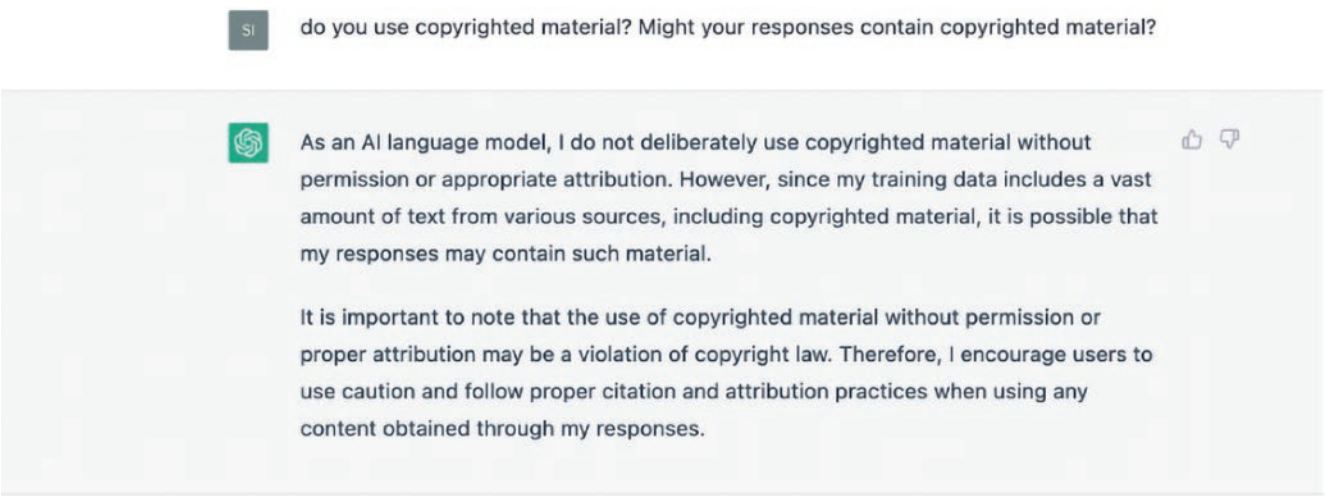


Figure 1. ChatGPT is happy to admit that it uses copyrighted materials and that it might provide copyrighted material in its answers.

Text creation for personal purposes

ChatGPT is adept at creating a well-turned newsletter to be included in a Christmas card on the basis of a few short prompts, or generating thank you letters based on the occasion, the gift received and a few short items of news. However, this does seem to rather defeat the point of this kind of personal communication, but just as for business communications, it should be acknowledged that written communication is hard for some people. In personal contexts, this can lead to individuals feeling excluded from social channels that require this skill. If it is acceptable for them to use a spell checker or a grammar checker to improve the text that they send to their friends, or to just type and print a letter rather than write it out in hard-to-read cursive, then why should they not use a content creation assistant as well?

However, some people are beginning to use ChatGPT in more creative ways, for example to create their profiles and responses on dating sites. The SWAN team (which runs a dating site so is probably quite invested in this) has pointed out that there are many potential harms that could be created by using ChatGPT for dating, such as reinforcing particular modes of speech and the development of false expectations and beliefs about another person, which can 'create emotional distress when the truth is revealed'.

Software development

ChatGPT is being widely used to help software developers. For example, ChatGPT is reported to have been used to create incident reports for software infrastructures and to create text for requirements documents. Other uses include using it to write the text for sprint stories and epic descriptions. These use cases free the developer or team lead from tedious text creation and editing, and require limited disclosure of proprietary information to ChatGPT. Whilst in many projects this might be feasible, there will be contexts where even this level of information leak is just unacceptable.

Additional use cases that do require extensive disclosure of code include commenting code, writing test cases, styling code or getting ChatGPT to explain someone else's code to you! All of these applications seem likely to boost developer productivity and free developers from the tedious hours of grind that often define their roles.

The extreme evolution of this is the idea of having a Large Language Model (LLM) as a complete system and just 'prompting' it to act in the right way. Essentially, you describe the programme that you want to implement in natural language, then ask the LLM to act as that programme.

This can certainly be done (I have made it act as a fizzbuzz programme), but there are limitations (I can't make it reliably do things that are not fizzbuzz). See figure 2 for a screenshot of a ChatGPT-backed API for fizzbuzz in action.

ChatGPT makes an excellent coding assistant. But there are problems with using it this way. It is unclear how much of ChatGPT (like copilot) is trained on open source code and the Free Software Foundation has raised concerns about this. So, any code that it generates might end up being interpreted as derived from proprietary sources, potentially creating copyright infringements.

The owner of a code training/tutorial website called 'Happy Coding' has convincingly demonstrated ChatGPT reproducing his code. Therefore it is quite possible that any output from ChatGPT contains copyrighted material and that material may not be appropriately licensed for the how you intend your

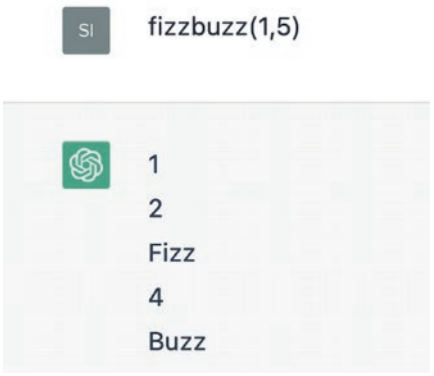


Figure 2: With a little prodding, ChatGPT will act as a Fizzbuzz API.

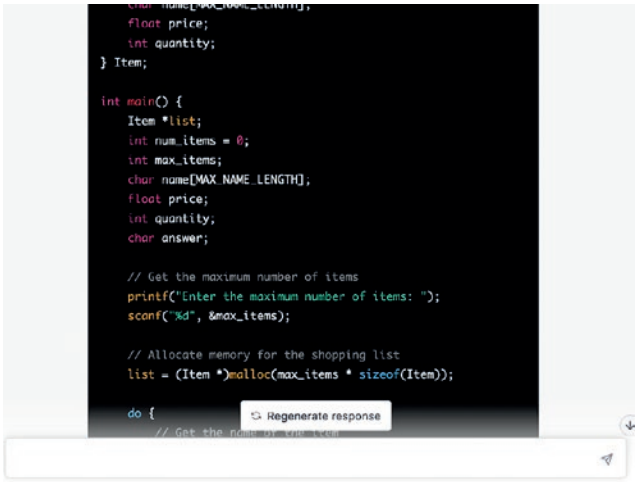


Figure 3: Some 'c' code generated by ChatGPT. Note that max_items could be any int value, potentially allowing a very large value to be given.

'new' code to be used. This could put your customers and you into a difficult legal position in the future, so caution is advised.

Another issue is that code generated by ChatGPT may be valid, solve a problem and run nicely, but may also have security vulnerabilities. For an example, see Figure 3.

The code in figure 3 is similar to a snippet of output created by Github Co-pilot (a competitor LLM) analysed in a paper by Pearce, Ahmad et-al that systematically assessed the security of code created by the Co-pilot model [5]. The issue is that a very large lump of memory might be allocated by the programme, potentially causing issues for other programmes running on the machine or even causing uncontrolled access to the memory of the machine.

One particular concern highlighted in [5] is that the code generator is trained on a repository of code examples that were developed over a period of time. Therefore, obsolete code, or code developed before particular exploits were developed or vulnerabilities were known is perhaps more likely to be in the machine's training set than modern secure code. It is also often the case that those writing text books include examples of what not to do in their tutorial material.

It is unclear that the training regime for a LLM like ChatGPT will reliably differentiate between examples of best practice and examples of incompetence!

Text creation by students

Every parent will be able to list the reasons why students should not use ChatGPT to do their homework, yet students everywhere are seemingly doing exactly that! In the past, students have already engaged ghost-writing services to cheat on assignments. By doing this, they have been able to gain advantages in return for taking some risk (of detection) and of course paying a significant fee. ChatGPT is free, and it is very hard to distinguish its output from a human's—especially if it is prompted in a skilful way.

An article in the Times Higher Education Campus describes this issue in detail and outlines some of the harms that this is creating.

- Inclusive and authentic methods of assessing students could be discarded because ChatGPT (and its successors) will be able to thwart them
- Radical shifts in curricula, teaching and assessment methods are needed to accommodate the new technology, diverting focus and resources from current efforts.

In addition, it seems likely that trying to handle ChatGPT-generated material will waste teaching and assessment time, frustrating and demoralising educators. This is part of a wider intellectual 'denial of service' attack that LLMs are facilitating which is discussed below.

Disinformation

Large-scale 'botting' attacks on social media platforms have been going on for a long time and have created a lot of damage and harm. We have also seen the web itself polluted by 'content farms' that are constantly engaged in an arms race with Google (other search engines do exist, but are generally irrelevant) to attract clicks by posting content that is highly ranked, even if it is misleading.

A tool like ChatGPT potentially opens a new era of disinformation. It is all too easy to imagine the Wikipedia editors being overwhelmed by plausible, seemingly human-generated edits. From an article in The Slate:

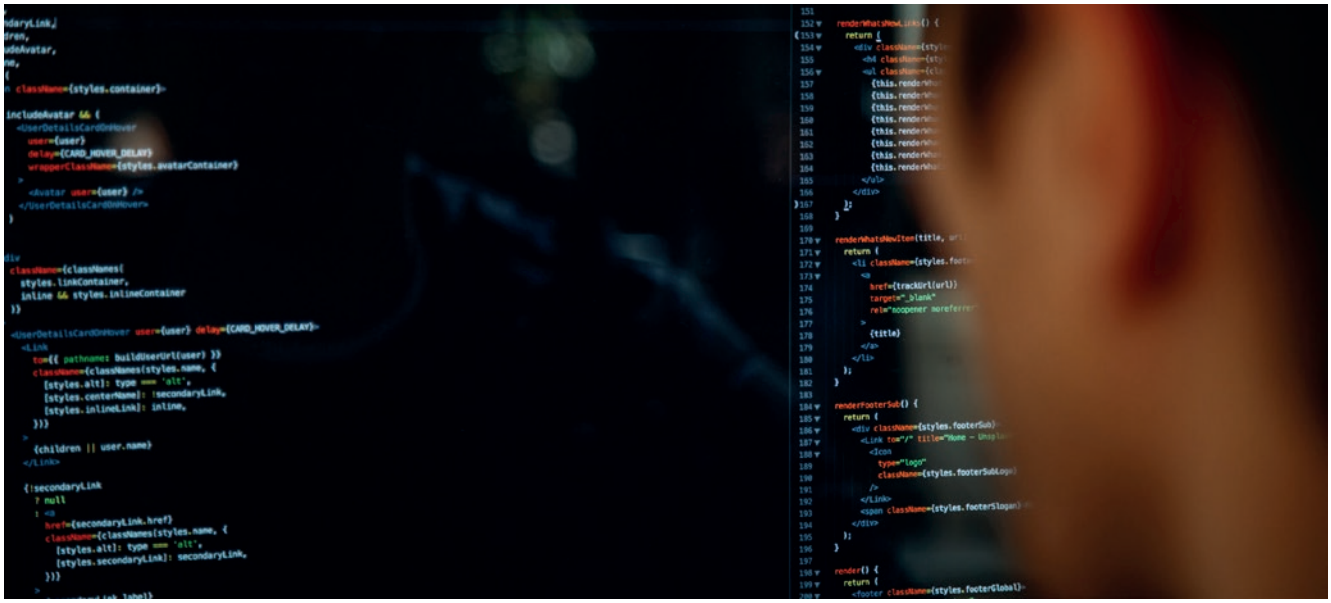
"The risks are that Wikipedia editors will find it more difficult to patrol content added to the site," said Heather Ford, an associate professor of digital and social media at the University of Technology, Sydney and the author of Writing the Revolution: Wikipedia and the Survival of. Facts in the Digital Age. "It will be easier for bad actors to create fictional content that is cited to yet more fictional sources."

Online doctorbots

In recent days , a series of tweets detailed an experiment by a start-up called KoKo that used ChatGPT to provide support for people with mental health problems. This has created a lot of comment, with many people stating that they find the approach and application to be unethical.

This is not the first time that internet technology has been used to experiment with users' mental health. In 2014, Facebook was discovered to have manipulated users' feeds to influence their moods. There was a lot of agonising about ethical review then as well. So, we have been here before and apparently some folks have learned nothing.

It is clearly very dangerous to provide advice about healthcare unless it is backed with a professional capability to dispense



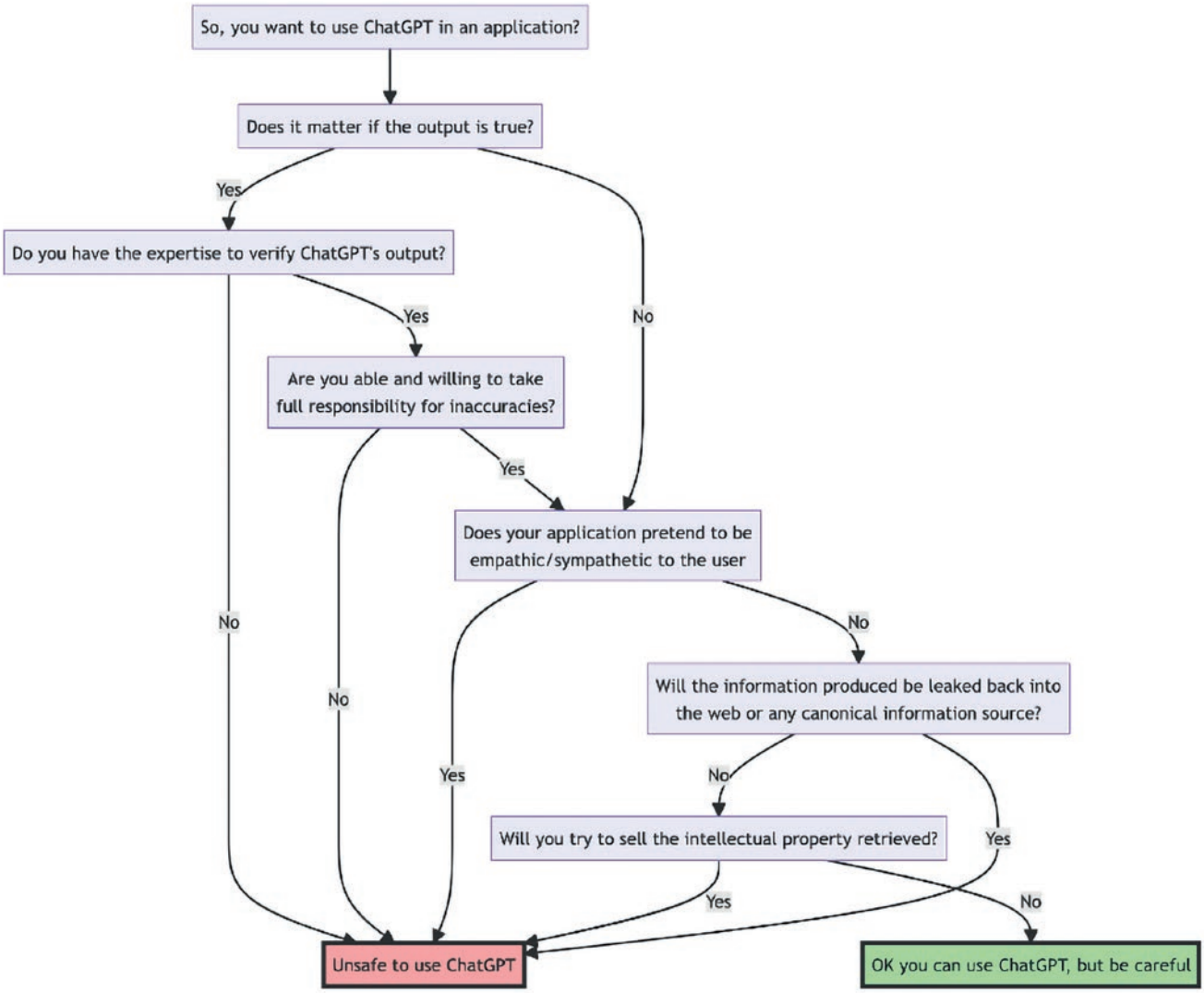
that advice. ChatGPT is not human and it has not led a human life - or anything remotely like it. Therefore, it literally can not develop the empathy or physical insight that a doctor has for a patient. Given that, ChatGPT can not perform the functions that a human doctor performs, but worse than that, ChatGPT can not provide authoritative medical information. By definition, ChatGPT's query matching process is approximate and probabilistic, and where approximate answers that are probably correct are good enough, this is sufficient. Where concrete facts are required, it can be very limited.

In some applications, the cost of a 'nearly right' answer is very low. Who cares if I trash my mower by following bad ChatGPT advice? I am just as likely to do the same damage looking things up on Google and using YouTube videos in any case, although validating the expertise behind them can be facilitated by looking for clues about who is presenting the advice. In medical applications, the cost is clearly extreme. Many other applications exist with similar costs - a bridge built using ChatGPT's instructions and ideas would not be a bridge I would be keen to cross.

What's off limits?

It is pretty clear that we are not going to put ChatGPT back in the bottle. The techniques used to create it are well-known, and although the amount of compute required seems to be heroic now, in the relatively near future it will be much more widely accessible. Even if compute prices do not shift down radically in the near future, the kind of compute required to create GPT3.5 is already available to many state actors, and a wide range of non-state actors. Google has announced that it will field a chatbot called 'Bard' based on its LAMDA technology which is so compelling that one internal engineer became convinced

Figure 4: A candidate decision tree for deciding whether or not to use ChatGPT for a task



it had a soul and Deepmind has developed a Chatbot called 'Sparrow' which is claimed to be technically superior to ChatGPT [2].

The big dangers are not likely to come from sophisticated super companies like Alphabet. Smaller companies with a 'move fast and break things' attitude are likely to be creative and adventurous with their application ideas. But very real harms to very real people are possible with this kind of system, and these can be easily and quickly implemented by small non- expert teams (as the Koko example demonstrates).

To be specific, there are at least four mechanisms of harm that LLMs such as ChatGPT can create for their users.

- Firstly, they can be untruthful or inaccurate; they can 'hallucinate' false answers to questions convincingly and mislead users into making poor decisions
- Secondly, the LLM may 'steal' the answer from a proprietary source, thereby potentially harming the creator of the intellectual property and the new user of the information
- The third type of harm is a breach of trust. The LLM may create an impression of empathy, concern or affinity that is simply false, and by 'lying' in this way may cause emotional distress
- Finally, using a LLM can distort and pollute information sources whether intentionally or accidentally.

Therefore, given that there are applications of ChatGPT that can cause harm, and given that we can identify the causes of these harms, it seems logical that we should be able to figure out when ChatGPT can be used (or when it should not). Already, people have constructed decision trees based on their insights (see the example at Figure 4).

Tiulkanov's chart shown in Figure 4 focuses on only one of the easily identified sources of harm for a ChatGPT-powered application, the trustworthiness of the output. When all of the potential harms are considered, perhaps Figure 5 is a more realistic approach.

Making things safer

Even though there are many paths to 'no' and only one to 'yes' on the more detailed decision tree in Figure 5, there will still be a lot of applications that get qualified as reasonable. But this will not make them safe. In order to have confidence in a ChatGPT- powered application, it is also suggested that the following steps are implemented.

1. There should be no deception about what it is that users are interacting with. You cannot give informed consent if you are not informed. Saleema Amershi et-al [3] have published excellent guidelines for interaction for AI systems. Importantly, these provide structure for considering interaction throughout the lifecycle of a user interaction. The guidelines cover how to make it clear to the user what they are interacting with and how to instruct them about what is expected of them. Amershi's guidance extends throughout the interaction, managing failure and overtime as the system becomes 'business as usual'
2. Users should have the option to not interact with the system. A real option - for example an alternative contact channel
3. There should be an impact assessment attached to every application. Put it on the website as you would a robots. txt file, or as you would add a licence to your source code. The Canadian AIA process offers a model for this sort of thing, but some fundamental questions are a good start. Who will it hurt if it works as intended? Who will be hurt if the chatbot goes wrong? Can anyone tell if the chatbot is going wrong, and can they stop it and repair the situation if it is?
4. If your system could have an adverse effect on others then there should be monitoring and logging of what the system is doing and how it is behaving. These should be maintained in such a way as to allow forensic investigation of the behaviour of the system, if required
5. If you are not personally and directly responsible for the system, a clearly documented governance process should be developed and maintained. Part of this should describe how users can call for help, and how they can complain about the system. It should also describe what the processes around addressing user distress and complaints should be.

In this paper, I have laid out what the potential problems are with ChatGPT- based applications and some tactics for avoiding and mitigating them in practice. I hope that our community deepens and develops these approaches in the near future.

With the correct controls and processes, new Large Language Models such as ChapGPT will provide great value in many use-cases, albeit with the essential controls and checks in place, to ensure users and end-users are protected from any misunderstanding.

1. Luby J, Kertz S. Increasing Suicide Rates in Early Adolescent Girls in the United States and the Equalization of Sex Disparity in Suicide: The Need to Investigate the Role of Social Media. JAMA Netw Open. 2019;2(5). doi:10.1001/ jamanetworkopen.2019.3916
2. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 'Bert: Pre- training of deep bidirectional transformers for language understanding.' arXiv preprint arXiv:1810.04805 (2018).
3. Glaese, Amelia, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh et al. 'Improving alignment of dialogue agents via targeted human judgements.' arXiv preprint arXiv:2209.14375 (2022).
4. Amershi, Saleema. 'Guidelines for Human-AI Interaction.' CHI conference on human factors in computing systems. CHI, 2019. 1–13.
5. Pearce, Hammond, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 'Asleep at the keyboard? assessing the security of github copilot's code contributions.' In 2022 IEEE Symposium on Security and Privacy (SP), pp. 754- 768. IEEE, 2022.