It's not AI that is the problem. It is humanity.

Mar 5, 2024

Leonie Raijmakers

Society needs to fix its integrity, social orientation and laws, including implementing accountability, to build and use AI safely. AI will amplify human traits without the emotional brakes normal humans have of fear, empathy or respect. The newest attempt of regulations by the EU are at least a start.

With the launch of ChatGPT and the news of Microsoft's first quantum computer being in serious development, AI has headlined fascination and concern across the globe in 2023.

The lack of good language models has been a barrier to many automation processes, not least of all the ever-increasing chatbots that are widely used: from health applications to banking advice. Though the limitations, multi-fold misunderstanding and inadequacies of these chatbots were annoying, there is something comforting in having an electronic system stumble, make obvious mistakes and need constant guidance. Even if it was just the chance to still feel superior as a human. Chatbot interactions would generally lead to either abandonment of the conversation to see if better information could be found elsewhere, or being referred to a human on the other side — as chatbots functioned less as a replacement and more as a support.

It has been shown that chatbots, in all of their inadequacy, clearly make a difference. Even in their unpolished form, chatbots had enough use and impact to support their widespread adoption. Including in the spheres of public health, such as the WHO applications for Covid19 and vaccine information, and manifold digital mental health support options. In multiple studies on digitised mental health support providers, automated bots appeared to give a feeling of comfort that something listened and could provide informed responses, but could not judge (e.g. Vaidyam et al, 2019). Seemingly. As all bots will have gone through training to do exactly that — judge — though logically instead of emotionally.

The availability of a highly developed language model, like ChatGPT, has rapidly improved such chatbots, and the entire digital communication landscape. An improvement that is making many human jobs redundant. That the impact on jobs reaches highly skilled workers should not be a surprise — anything analytical, or even creative, that is mostly done on a computer is ripe for takeover; such things are much more easily automated than a robot that will automatically fold and hang your clothes with minimal space invasion… So, the fear of job loss has been present in largely the wrong sectors — though I do genuinely hope the lady asking whether people have no liquids in their bags before entering the scanner section at Heathrow will soon find a more fun and interesting job for herself. Both as I wish her much better than dealing with stressed and grumpy travellers, and as I hope the London airports upgrade to a system similar to Schiphol, as I have had enough of having to remove my laptop from my bag for over 20 years (who calls this rapid development?).

But let's not get ahead of ourselves, and first look at two big questions in AI that cover the general fears towards it. Firstly: who becomes responsible for the mistakes AI makes? Secondly: will AI go out of control and start leading a life of its own? These two questions are quite obviously related, go deep, but aren't that difficult to answer theoretically. *The difficulty lies in application and practice.*

To address the first: it's pretty simple. Like any tool you use, responsibility lies with how it has been created and how it is used. If someone crashes their Toyota (we're still speaking normal, manual driving here, but technically the same applies to a self-driving car), the crash is in first instance their fault. Should the cause lie with a problem with the car — e.g. the engine overheated due to a technical issue, leading to loss of control — it is the responsibility of the manufacturer, i.e. Toyota. If you use your car in a heist, you are responsible for the heist, not the car and not the manufacturer. However, if the car has a special function built in that facilitates heists in particular and isn't necessary for normal, driving functions, one could start questioning the manufacturer for providing such a function.

Though chatbots are only a limited application of Machine Learning and AI compared to all that is and will be impacted, they are a usefully simple illustration for some of the questions regarding AI and responsibility. For example, the WHO covid19 chatbot — which provided information about covid19 infections globally — had a strong political orientation. Returning national covid19 statistics on request by texting a country's name, the chatbot functioned fine with France, Germany, Thailand, China. Faced with queries for Taiwan or Hong Kong however, it malfunctioned. 'Hong Kong' gave 'Hello'. 'Taiwan' gave 'thank you. Is there

anything else I can help you with?' or something along those lines. This strong political orientation could be sniggered at and put aside as incompetence. Human? Machine? A combination of both? Everyone knows it's the former dictating this 'error' — and an easy workaround would have been to provide information on the basis of 'states' or 'regions' as well as countries. Such more localised information would not only have provided a political work-around, but been useful too, as regional information is extremely important for informing healthcare decisions in the middle of a pandemic. Refusing to provide it on the basis of politics shows up flawed priorities that have nothing to do with health. Imagine that such a set of flawed priorities is the basis of training AI machines on how to modulate their responses. AI is a very smart child, learning from its parents even when the parents do not realise. 'Where did they gain that swearword from?' is one thing. 'Where did this machine influencing government decisions get their advisory orientation from?' is something entirely different on a global impact scale. The competency of language models is now less a system problem — as the system problems get fixed, it is increasingly a human one. This is where tech really needs to start incorporating philosophy and law, and should have long ago, as signalled throughout the early 2000s with the discussions on the value of arts and humanities vs the large interest and investment in science; and continuous conclusions that the latter needed the former.

Examples with pharmaceuticals might also be distinctly useful, as there is significant interest in and development of AI applications in healthcare and population health. To properly prescribe medication to patients, you need trained medics and pharmacists. Mistakes in medication can be the patient's: for not adhering to their prescription, self-medicating with the prescribed medication, or taking other medication or products that interfere with the workings of the medication — of course this is assuming the patient has been properly informed of the consequences and what to avoid. The mistake can be the doctor's for incorrect prescription or not informing patients properly; or the pharmacists for not checking counterindications.

Yet of course, before all this, it is the responsibility of the pharma company to ensure its products work, through development and testing, and are manufactured properly. On top of that, it is the responsibility of national and international drug safety boards, such as the FDA, to ensure this is the case and only drugs that are safe to use are approved for their specified indications. Medically oriented AI in first instance needs similar checks and approvals before entering the market — or should. Though there are some FDA regulations regarding medically focused applications, these are by no means significant and definitely not adequate.

It is a space where international regulations are lacking. This means opportunistic experimentation abounds, and raises — or at least should raise — questions of integrity and whether proper medical ethics is appropriately embedded when medics aren't part of the development and business teams. Which is often the case.

Let's head back to something a little simpler. Like any tool, AI is open to being abused for nefarious means, as much as it is used for good. Including, for example, the production and sale of child pornography images with AI image generators (BBC, 28 June 2023). The balance of our fear, rightfully so, is that the better it gets at being used for good, the more efficient it is in being used for bad. This is where regulations — and not just regulations but also active penalising — need to come in, on both sides of the responsibility spectrum: developers and users. It is also where AI researchers and developers are (rightfully) concerned.

Several highly educated people around me have claimed that academics and other highly educated people are averse to chatbots. However, even they tend to use chatbots — and are likely the biggest users of ChatGPT, a sophisticated chatbot. Believing you, and others with you, won't fall for increasingly sophisticated automated models of persuasion isn't just unrealistic and unlikely: it is highly dangerous.

At the risk of giving bad people bad ideas, under the presumption that this is already in use, realising that people consistently fall for scam emails that are less than sophisticated, one can only imagine what happens once sophistication has been handed on a silver platter to scammers — completely automated at that. To the general public this is a much greater, and an already existing, risk than dystopian fears of hostile missile takeovers, yet gets little attention.

That ChatGPT is better at answering medical questions than basically trained volunteers — and increasingly even than some physicians — appeals to people. What makes this scary is not the AI aspect but the human aspect, where sensationalising research for news purposes could lead to people using ChatGPT exclusively to answer medical questions instead of going to see a doctor or a specialist. In this case, fault lies not with the AI or the researchers, but with journalism. Another group of responsible parties to be aware of, as news platforms (including popular social media channels), should be held responsible for the consequences of improper reporting. Journalism has a (fantastic and respectful) function to keep others and societies accountable, yet at the same time is to be held accountable for the consequences of incorrect reporting. In a way, news articles are often the manual to understanding the world, including

in this case how AI can be used, and in contrast to most manuals (hello IKEA, Philipps and Samsung), these ones are actually widely read.

Further dangers with incorrect reporting is the increased abuse of (social) media channels to spread disinformation. Efficient and smart producers of text, and now also visualisations, like ChatGPT, support the increased abuse of information channels to influence people. Manipulation of people through social media channels was already able to influence elections and national voting behaviours from at least 10 years ago, with smart but less efficient systems. The fear that efficiency in creating increasingly real-life presentations — combined with psychological impact optimisation — escalates such possibilities is very real, and means AI enables efficient weaponisation of emotional manipulation. This probability would have been present in the minds of anyone deciding to work with AI and psychology — including myself when taking up a research position in Hong Kong that looked at automisation of persuasion models, particularly focused on public health messaging and vaccine hesitancy. When taking up such work, some will have set themselves lines that should not be crossed, and stepped away when they were; some will not — for a variety of good and bad reasons.

Journalists and medical experts have pointed out in articles that there is a special limit to ChatGPT — one where humans will (always?) have an advantage. Drs Dranove and Garthwaite, Bart, and journalists Beerends and van der Ster write that AI, with its focus on efficiency, lacks and will lack the empathy and understanding humans have. Personally, I do not believe empathy is a threshold AI cannot cross. Feeling empathy? Yes, that is unlikely. Simulating empathy? Absolutely. Let's be honest. Plenty of people feel empathy without being capable of putting it into words. Conversely, one does not need to feel empathy to be able to communicate it. Think of psychopaths[1], and how they can manipulate people without an ounce of emotion. Although often certain quirks of their humanity may betray them — the fear of being discovered or joy of getting away with something. Alternatively, their lies might build up to become increasingly visible to those who have started to have suspicions, simply through one lie not adding up with another.

However, if you translate psychopathic skills into an emotionless AI, this becomes an optimising system that has learned (and keeps on learning) to manipulate people whilst devoid of emotion, utilising a web of lies based on all of humanity's online knowledge — i.e. so deep as to be impossible to discover, and unlike the Titan at great depths, without a detectable structure that implodes under pressure. In fact, it will have an optimised structure of flexibility to learn and adapt even further. With such constant training to hide lies

increasingly well — which is something already observed within ChatGPT[2] — AI does become an extremely dangerous tool that goes well beyond the manipulation algorithms of Facebook influencing people's choices, voices and views. The combination becomes, well, unthinkably powerful and uncontrollable.

This danger becomes more pronounced when such a system gets integrated into all aspects of life and physical existence: for example China with their increasingly close integration of biodata within daily operations systems; such as local transport, shopping and healthcare [this is not for the sake of giving China as a bad example. It is an example of a highly developed, integrated system, that many other nations are to some degree moving towards]. Having a by default greedy, narcissistic psychopath of an electronic system ruling our world is extremely unlikely to make the world a better place in a way that is equitable. Instead, it will likely make the abusive systems already in place increasingly efficient at hoarding affluence towards a small group of people.

Of course, for all their bad rep, psychopaths aren't necessarily evil. Whether they are or aren't depends more on their other interests and characteristics. The orientation of a narcissistic psychopath will depend on what feeds their narcissism. When this aspect is aligned with the good of society, this will hopefully mean the system is generally non-malignant. As Taylor Swift sings, 'covert narcissism I disguise as altruism' is very much a thing, and always has been. It is hard to complain about narcissistic altruism when there is a large number of good people benefiting from it in a sustainable and helpful way — even if its premise is based on feeding the less morally-oriented 'ego' of a person or system or two.

Hopefully most AI systems will and can be regulated enough to be benign. 'Hopefully', as even if programmers mean good, AI is a technological system that goes so deep now that much of how it functions is unknown, and therefore uncontrollable. Even what is known and controlled needs to be consistently corrected — with increasing effort and sophistication, even though it can be braindead work. Here, human error once again creeps in — and not on the sidelines but in the centre of the system. *Hiring smart people to do the braindead work of catching sophisticated errors is not going to be something businesses are interested in.* For if it is only smart people catching it, isn't the product generally fine to be used? The current concerns about the spread of mis- and disinformation would be infinitesimilised (annihilated?) as soon as this becomes an accepted stance for a seemingly 'high performing' or even 'near perfect' large language model AI product. *That human laziness and greed lead to exacerbated complications within an exacerbated system based on human systems of*

laziness and greed is a profound given. Circular logic? Absolutely. For that is also what is feeding your language model.

The discussion in the paragraph above of the mistakes AI will make — and already does — is of course only touching on basic human failing, errors, a bit of laziness, a little greed. However, if the narcissistic psychopathic system has programmers, developers using it for their own greedy ends (or of their bosses'), this is what will inform and teach the system. In short, it is extremely vulnerable to any loss of integrity, alertness, or quality in the people developing it. Even with all those in place, it is already outsmarting developers and users, making it uncontrollable. As a system of which the workings are unknown, and what it has taught itself is a mystery.

Yet the largest concern lies not in the workings of the system itself. The largest concern of optimised AI tools are the inherent flaws in humans. The more efficiently you copy a human's capability the more effectively any flaws get integrated, with the risk that these flaws come with benefits, exacerbating the traits in an unmonitorable system. Just like online social media systems need constant monitoring and redirection — often with underpaid and undertrained human monitors getting exposed to the most gruesome parts of society — an AI system will need constant steering and monitoring, but by highly expert monitors with high levels of integrity as well as a high incentive to ensure both. This work doesn't just include regulating answers to queries that are made, i.e. the uploads, but in this case also regarding the outputs and any signs of behaviour (model strategy) learned that is or will likely lead to corrupted outputs. A lack of high-quality monitoring will lead to increasingly embedded high-quality 'hallucinations' of AI programmes.

The question is whether a system that has been trained on outputs that are by default flawed (human outputs) can be steered to become "clean", or whether it will learn to become so efficiently dirty that it is hard to discover it isn't clean. And let's be honest. Often humans do not even recognise this within others, let alone themselves.

Where we get lazy or inert, the smart with initiative will take over. Humans have always been creative in making systems efficiently cater towards their interests. Trying to stop this and creatively make systems fair and equal is something large numbers of us are inclined to and to some degree capable of. However, ensuring this, means ensuring societies remain focused on rewarding fairness and integrity. Or, to be more realistic, *become* focused on rewarding fairness and integrity. Something the current system is increasingly leaky of across the board,

including not just governmental systems and business, but also researchers, and respected global academic institutions (e.g. two Professors at Harvard in 2023 — one expert on honesty who was accused of data manipulation, and another who was possibly removed from Harvard due to economic interests of the university towards Facebook influencing their decision to remove a disinformation expert who shone a light on Facebook's economic interest in disinformation; but many more reported and unreported cases exist — tbc). Where simple integrity policies are included as a box-ticking exercise, but the box ticking does not include actually applying them — hopes of having AI systems that are continuously, stably and reliably fair, equal and highly integrous are pretty slim.

[1] Gianluca Mauro, from AI Academy, posted a similar observation of AI systems being comparable to psychopaths.

[2] Looking into references requested of ChatGPT showed either marginally useful references, or ones that were entirely made up — these references did not exist. This indicates a 'lazy' and uncorrected learning model, where the idea of giving references has been learned, but not the ability of giving actual, proper references.