ChatGPT – a catalyst for what kind of future?

Statement of the Digital Humanism Initiative, March 2023

The release of ChatGPT has stirred worldwide enthusiasm as well as anxieties. It has triggered popular awareness of the far-reaching potential impact of the latest generative AI, which ranges from numerous beneficial uses to worrisome concerns for our open democratic societies and the lives of citizens.

This development offers an unexpected, but welcome, occasion to *explain* to the wider public and policy-makers what AI tools like ChatGPT are and how they work; to *highlight* beneficial uses, but also to *raise concerns* about its considerable risks, especially for liberal democracies; to *underline* the urgency for public discussion, society-wide response and the timely development of appropriate regulation; and to *argue* that academic research needs a fair chance to set research directions independently of the large corporations and their huge investments that are now being poured into further commercial exploitation.

ChatGPT: some basic facts

Released by the Open AI company near the end of 2022 as a conversational version of their Generative AI models, ChatGPT (Generative Pre-Trained Transformer) is a Large Language Model (LLM) based on the innovative combination of both unsupervised training and reinforcement learning from humans[1]. The large data sets come from sources such as books, (news) articles, websites or posts, or comments from social networks to perform its core function as a dialogue system simulating human conversation. It achieves this by ‘estimating‘ via probabilities which word(s) is likely to follow the previous word(s). This is done in accordance with specific writing styles or tones which, in turn, creates the illusion of conversing with a human. While apparently good at mastering this aspect of language, these systems are rather limited at the functional level, e.g., lack of reasoning and abstraction capabilities, or very limited situation modelling. We humans speak to communicate with other humans with the intention of achieving some goal. We tend to attribute this intention to all agents that produce language. We are thus easily seduced to project human intelligence as we understand it onto machines capable of some form of language imitation.

It is important to underline that such models can build convincing combinations of words and sentences, but it does not have human understanding of our questions nor its own answers. Neither does it have an understanding of what ‘facts‘ are and is prone to produce factual errors and to ‘hallucinate’. Open AI admits the limitations of its models: ‘ChatGPT sometimes writes answers that sound reasonable, but in fact are incorrect or nonsense’. This has been the

[1] On March 14 a successor, GPT-4, was introduced, which accepts both image and text inputs.

reason for characterizations such as ‘stochastic parrot’ or, less kindly, ‘confident bullshitter’ and others which, of course, are occasionally applicable to humans as well.

OpenAI and other companies have now entered a fierce competitive race and continue to gather feedback data from a rapidly increasing number of users. This means that all of us are the subjects of a huge ongoing field experiment that these companies are conducting - without our consent.

Potential ‘good’ uses

A common attribute of potential uses that are deemed to be beneficial is that AI tools like ChatGPT are used as ‘side-kicks’ or assistants to humans, complementing humans ability to cooperate and participate in society. Such an approach, augmenting instead of replacing humans, has already been argued in the early days of AI. Under this general proviso, there is a long list of potential beneficial applications ranging from assistance in the preparation of legal briefs, translation, programming, chip design, material science to drug discovery and, of course, education and training. This can stimulate innovation, lead to new business opportunities, and create productivity gains in many sectors of the economy.

The history of technology demonstrates that many uses cannot be foreseen, as the social, economic or cultural contexts in which new technologies are adopted and appropriated by users vary considerably. Users, therefore, are not merely passive consumers, but have shown in the past the ability to twist or invent new uses better suited to their needs.

Potential ‘bad’ uses and risks

Currently, the list of concerns about potential abuses of the technology is long, including

- ‘Industrial level’ production of ever more convincing scam emails of all kinds and their wide diffusion by a range of actors, including some governments;
- automatic production of fake news and large numbers of websites for targeted disinformation campaigns by individuals, businesses, and states;
- automatization of communication with scam victims e.g. instructing them how to pay the ransom asked;
- fast, efficient production of customised malware code including new breeds of malware that can “listen-in” on the victim’s attempts to counter it;
- deep fakes by systems trained on images.

Of concern is also the risk posed by the use of these systems by young people during their formative years in school. This is the period where key human cognitive capabilities are developed. We are worried that excessive use of these tools as (themselves-not-well-understood) shortcuts to learning and practising could severely impair these capabilities.

We are also concerned about the lack of transparency and accountability coupled with the loss of shared reference to what is true/false or good/bad. There are reasons to worry that models may intentionally be misused or, even inadvertently, cause dramatic accidents and social disruption that leave completely open who can be held responsible for the harm and damage. It is also very likely that a new ‘arms race’ between ‘robbers’ and ‘cops’ will be set into motion, with cybersecurity experts constantly having to upgrade preventive measures. This comes at a high economic cost but might also lead to further curtailment of civil liberties.

Finally, we are concerned about the enormous concentration of power, resources, and prioritisation of future AI R&D directions in the hands of Big Tech, if the current unconstrained development continues. There are sufficient historic examples to show that the concentration of economic power rapidly leads to a concentration of political power and vice versa.

ChatGPT, Big Tech and liberal democracies

During the last decade numerous studies have shown the fragile state of liberal democracies around the world, concluding that they are ‘back-sliding’ or even in ‘precipitous decline’. There is also a geographical retreat as by now over half of the world lives under authoritarian regimes. Economic inequalities, the effects of unrestrained globalization and constitutional fault lines are among the leading causes for the decline. These are closely intertwined with the role played by Big Tech and their platforms.

The concentration of economic and political power in the hands of a small elite heading a small number of big companies is a major concern related to the outsized influence they exert on democratic processes, institutions and the erosion of the public sphere. Their political power, besides lobbying, stems from the increased capabilities made available via their platforms to nudge, herd, manipulate and polarise public opinion. These capabilities have been and are being used by internal and external perpetrators who seek to undermine democratic processes.

It is in such a context, evolving before our eyes, that the threats to liberal democracy may get exacerbated by the AI race among these tech giants, kicked-off by the release and phenomenal publicity of ChatGPT. Huge investments have poured into a rapidly evolving digital ecosystem whose direction, scale and further investment is solely determined by a few companies.

The cost of training a very large AI system like ChatGPT and the associated requirements for computing power and data sets, represent a real danger as the power of these tools is concentrated in the privileged hands of a few companies and a few governments. Nobody can build such tools in a garage and academic institutions are less and less able to keep up with these companies and the generously funded start-ups that are then acquired.

If the vicious cycle of the platform economy with its ‘winner takes it all’ phenomena is allowed to continue, any remnant of the perhaps idealistic vision of a pluralistic digital ecosystem developing AI tools that empower citizens, complement and thus augment their capabilities to participate meaningfully in an open democratic society – the digital humanism vision to put it shortly – will be brutally swept away.

In summary: The development of AI tools like ChatGPT, that questions what it means to be human, cannot be left in the hands of a few powerful companies. Such AI must be a public social good and democratically governed.

What needs to be done

The centralised control of this experiment and the related decisions on AI research directions represent a threat to the sustainability of liberal democracy which is clear, imminent and vividly highlighted by the glamour and publicity currently surrounding ChatGPT.

The fact that this threat is raising flags of concern among political decision makers (note the recent turmoil in the European Parliament debating the AI Act triggered by ChatGPT), academic networks, initiatives like Digital Humanism, and other like-minded ones, gives hope and motivation to take action during this fortuitous but probably limited window of time.

The need for regulation and our concern that unregulated AI will, on the whole, be bad AI, have not gone unnoticed. In the European Union, the EU AI Act is under intense discussion with the aim to be approved in the coming months. In the USA there have been several antitrust suits for monopolistic behaviour by the Federal Trade Commission and the Department of Justice, with the latest one filed against Google on January 25, 2023. Related to this, in the transatlantic Trade and Technology Council, the USA and EU pursue a Joint Roadmap for Trustworthy AI and Risk Management. The implementation of AI policy must be continuously monitored and updated in a dynamic way.

What needs to be done, in addition to pressing for good regulation and its implementation, is to keep the general public and policymakers informed. They must be made aware of what is at stake regarding the future of democratic institutions and processes and the risk that citizens become pawns in a closed competitive race about profit and market shares. The public sphere for open deliberations and

participation is at risk of being taken over and flooded by content that is deliberately designed for misinformation, utter nonsense or undermining the sense of democratic, collective belonging.

We, the Digital Humanism Initiative, academics and researchers in the computer science domain and in the social sciences and humanities, working on current developments of AI and its societal and cultural impact from the perspective of Digital Humanism, feel responsible to inform and explain to the wider public and to policymakers the opportunities and risks that come with ChatGPT (and similar AI tools). Future directions of AI research and development should be driven by human-centered concerns and human needs, in a future which is not dominated by the profit-oriented goals of large companies.

We commit to apply the digital humanism approach in our AI research and development, remain publicly accountable, and stay open to constructive debate in order to improve our approach as generative AI technology, its uses, and our understanding continue to evolve.