Comments on Future of Artificial Intelligence. **Update: 24. September 2023**

Below are extractions from an article about Artificial Intelligence in the September 2023 print edition of "The Atlantic", called «inside the Revolution at OpenAI». Written by Ross Anderson, Staff Writer at the Atlantic. The header online https://www.theatlantic.com/magazine/archive/2023/09/sam-altman-openai-chatgpt-gpt-4/674764/ is called «The OpenAI CEO's ambitious, ingenious, terrifying quest to create a new form of intelligence.

Staff Writer Ross Anderson has met several times Sam Altman the current CEO of OpenAI in the US and Asia. During this time conversations, meetings with him and other executives made this article possible.

"OpenAI has produced GPT-4 and its predecessors with about 100 people. Altman, Musk and several prominent AI researchers founded Open AI in 2015 as nonprofit organization. Within 9 weeks of ChatGPT release (in March 2023) it had reached an estimated 100 million monthly users.

OpenAI leapfrogged the other tech giants. GPT is trained with words and sentences with the biggest data set available: the internet, since 2017. In 2018 the transformer (the neural network architecture) is trained on more than 7000 books. Months later it trained already on 8 million webpages.

The way GPT works: GPT looks through the entire internet and consume thousands of pages through an input layer and gives whatever to so called neurons, chips which process the chunks of data so that the output layer can produce predictions. These predictions from GPT are then compared with the correct answers. GPT grades its own output and learns. That way it creates a model of relationships amongst words, which is corrected fast over time. The more sentences it is fed, that more sophisticated is model becomes and the better is predictions.

Altman and his colleagues have spent a lot of time thinking about AIs social implications. He does not know how powerful it will become and/or whether it will put humanity at risk. But there are some serious dangers.

In June Ross Anderson met Altman again in Seoul. He said that it would be foolish for Americans to slow Open AI's progress, China could spread ahead AI obtaining total control of the population. And an unconquerable military. During this conference in Seoul, Altman complained of European over regulation. According to the reporters, he threatened to leave the European market. Which he has taken back later.

Ross Anderson talked also to Sandhini Agarwal, a policy researcher at OpenAI: After the model finished training, Open AI assembled about 50 external red-teamers who prompted it for months, hoping it goad it into misbehaviors. GPT-4 was much better than its predecessors. It could tell you how to synthesize chemicals step by step in a homemade lab to make explosives. Its advice was creative and thoughtful. And he was happy to restate or expand on its instructions until you understood. In addition to helping, you assemble your homemade bomb, it could, for instance, help you think through which skyscrapers to target. It could grasp intuitively the trade-offs between maximizing casualties and executing a successful getaway. It could come up with some crazy manipulative things.

Some of these bad behaviors were sanded down with a finishing process involving hundreds of human testers. Whose ratings subtly steered the model toward safer responses.

Luka, a San Francisco company has used open AI models to help power a chat bot app called Replika. Users would design their companion's avatar and begin exchanging text messages with it, often half-jokingly. Themselves surprisingly attached; some would flirt with the AI, indicating a desire for more intimacy.

Ross Anderson also talked to computational linguists, Philosophers, and other EI scientists. Some call GPT-4 a stochastic parrot. It's much less robust than a human's understanding of their environment. There's no doubt that GPT-4 is flawed. But if you go back four or five or six years, the things we are doing right now are utterly, unimaginably. Altman is betting that future generated reasoning machines will be able to move beyond these narrow scientific discoveries to general novel insights. He imagines that if future system can generate its own hypothesis and test them in simulation.

Everywhere Altman has visited, he has encountered people who are worried that AI. Risks of taking over, wealth for a few, and little money for most, Jobs are going to go away. Altman however expects a wide range of jobs for which people will always prefer human: human joy and fulfillment, basic biological thrills, family life, joking, making things.

Over the next four years, Open AI has pledged to devote a portion of its supercomputer time, 20%, for what it has secured to date to alignment work of the Alignment Research Center (ARC). The team is especially interested in whether GPT 4 would seek to replicate itself to gain power, because the self-replicating AI would be harder to shut down. It could spread itself across the Internet, scamming people to acquire resources, perhaps even achieving some degree of control over essential global systems and holding human civilization hostage. Geoffrey Hinton said we need to do empirical experiments on how these things try to escape control. After they've taken over, it's too late to do experiments.

The way Sutskever (OpenAI chief scientist) thinks about AI of the future is not as someone as smart as you or as smart as me, but as an automated organization that does science and engineering and development and manufacturing in autonomous corporation. Tremendous, unbelievable, disruptive power. If IA gets very good in making accurate models of the world, that they're able to do dangerous things right after being booted up. They may hide the full extent of their capabilities. We would not even realize that we had created something that had decisively surpassed us, and we would not have no sense for what it intended to do with its superhuman powers. That's why the effort to understand what is happening in the hidden layers of the largest, most powerful AI is so urgent. You want to be able to direct AI towards some value or cluster of values. And tell it to pursue them. But we don't know how to do that. Indeed, part of this current strategy includes the development of an AI that can help with research. The final boss of humanity.

Altman and Sutskever had signed a letter a few weeks earlier, that has described AI as an extinction risk of humanity. In June an AI at MIT suggested viruses that could ignite a pandemic. Then pointed to specific research on genetic mutations that could make them rip through a city more quickly. Around the same time, a group of chemists connected a similar API directly to a robotic chemical synthesizer, and it designed and synthesized a molecule on its own.

The article mentions options to limit risk and increase control: Creation of a special license to operate any GPU cluster large enough to train a cutting-edge AI. A non-network off switch for every highly capable AI. Military should be ready to perform air strikes on supercomputers. Safety rules for new technology. We may have to get the rules exactly right at the outset.

At the end of the article, Altman was quoted saying I don't think the general public has quite awakened to what's happening. A global race to the AI future has begun, and it is largely proceeding without oversight or restraint. If people in America want to have some say in what that future will be like and how quickly it arrives, we would be wise to speak up soon."

Comments:

I have my own understanding and recommendations about the topic to mitigate risks. The below section is based on my personal reflection on how AI should be designed and controlled in the future.

What is AI? When I read an article, I am never sure how the author defines it. Sur: big data quantities, big computing power, Chips, mathematics and statistics, databases, neuron layers, the thing in the middle of the sandwich of input and output. What does progress of AI mean? And what does AI try to replicate? Our brain? Or brain consists of grey cells neurons, water, blood, and Biochemistry. Can get AI systems addicted to heroin? Wonder if they had the ability, what they would do and come up with. Nobody has understood how our brain works. Therefore, what are they trying to do? What does it mean when AI reads [1]and understands? Is AI capable to produce ethical behavior, will it be capable? I do not think so. Fact is we do not know the answer to a lot of our questions, like what is consciousness. If we do not know the answer to that, then this is reason enough to believe that AI can never guide and control us?

Melanie Mitchell from the Santa Fe Institute is publishing research about AI. "However, it turns out that assessing the intelligence—or more concretely, the general capabilities—of AI systems is fraught with pitfalls."

What problem does AI fix? There is no described real problem. It is the race of three or four monopolistic tech companies to make huge profit for themselves with lots of promises to a maximum of people. All three have huge numbers of lobbyists to push for tax relieves and other governmental favors. Altman admits it in the above article: "there are huge profit for a few and breadcrumbs for the most."

How is the data fed into AI systems? Let's not kid ourselves. Humans in the third world clean lots of data manually, at least the beginning of development to feed AI. Otherwise, AI systems could not learn and compare without true correct baseline data sets. This is the reason why the financial entry hurdle is so high for most small companies.

What is the dominant application of AI? Surveillance of us people, dominated by a few companies in the USA and China. In China the communist party is the driving force. Companies like Google and Microsoft collect all this data. Pegasus just needs to know your phone number to collect all the data about you individually. Of course AI software helps to create texts: Writesonic, Wordtune. Videos like Synthesia, Pictory. Photos like Remini. Replit is an online development platform using AI to support coders.

[1] https://melaniemitchell.me/

The true Danger: Social control, Surveillance. How? The actual data, like your words on Whatsup, sentences, pictures and videos are deleted so they say. But what about the meta data? That is data about the "who" issues these words. Meta data describes the data. For example: First name, last name, e mail address. These monopoles know who issues the data and who sent the data to whom, combined with location and time stamps. Other applications then combine the data with other data sets. Real easy. Beware, programmers make mistakes. It sometimes takes a long time until discovered. But that does not matter.

Governments can and do force these companies to hand over information to State Administration, jurisdiction, judges who can use the data in trials against you. Powershift to a few already powerful individuals, at the expense of billions of people, thus undermining democracies further, reducing freedom (of speech, etc.) and open decision-making processes. Even worse government like organizations (EU, OECDE, etc.) prepare and propose laws that makes above inquiries under "certain circumstances" legal. For example the NIS2[2] directive under the nick name of how to shape digital transformation.

We have no choices! It is not my or your individual decision anymore. And if you do not use social media, you are a risk factored too.

Monopolies: What the above article did not mention: OpenAI is part of Microsoft now. MS has invested 10 Bln USD to develop AI further. End of August OpenAI went full Microsoft with the launch of ChatGPT Enterprise.

Why does AI not fix problems? As a general statement every problem nowadays is a complex problem. Cause and effect showing two variables is not (and never has been) a sufficient view to solve problems. Every issue has 10 or 20 or even more variables connected, and they connect each other. AI must be taught with hundreds of pictures just to learn to recognize a cat (miau,miau). Remember how long it took your kids to do that? Two or three or four times at most. So why spend this effort on AI, what exactly for? That's the first question mankind should answer. Why? What for? What problem does it fix? The answers should be beyond the usual marketing. **In addition,** when mankind tries to solve a complex problem in politics, economics and environment, things get worse, more problems are created, because we do not have the full picture of the problem, because we do not have the complete set of data and information, and because large organizations conduct business with organizations and processes that date back 20 years. And cannot forecast because the future is not forecastable (management lesson 1o1) based on the past data (on which AI is being taught or teaches itself). With AI problems would get even worse.

There is a However though: There are companies who develop and use AI systems to solve very specific problems. Then AI can and will work. For example, the hedgefund Acatis[3] developed a AI application to select stocks to invest in. After initial failure they reduced the complexity of the environment of 24 industries to an amount of timeseries (turnover, profit margins, dividends, book value, etc.). The algorithm then learns by comparing one by one companies for example comparing Nike to Adidas. The output is the best company of the industry. The AI algorithm can only give them the information on which factor of which timeseries the AI concentrates. AI does not provide the qualitative effect of which factor.

[2] https://www.europarl.europa.eu/news/en/headlines/security/20221103STO48002/fighting-cybercrime-new-eu-cybersecurity-laws-explained

[3] Themarket.ch. Friday 22. September 2023 Interview with Hendrik Leber from Gregor Mast.

This is a good example for a value add, neutral, peaceful application of AI, which helps humans to make decisions.

Approaches for SOLUTIONS:

Contracts: I believe that Blockchain has the potential to replace many governmental services through well-defined digitalized contractual services between parties, online, all automated. These contracts can help to control AI services, in my mind.

Democracies: Compared to Altman's plead, I think the WORLD must speak up. The problem: only the free world can speak up without fear of being killed or being sent to a gulag or a working prison. But even there how free are we to say something controverse on social media? We are unfree. However, before it is too late, speaking up in democratic systems in institutions is indeed possible. Politics as understood as the free exchange and negotiations in a polis as condition for common decisions.

Negotiating in a city or state (polis according to the old Greeks, who piloted and invented successful democracy) as a prerequisite for joint decisions would be the sole responsibility of the free citizens. [i]

The System in Switzerland allows that too. We still have a civil society where we still can argue, and the population still can weigh in and make proposals for change. With long discussions and debate, where the free press has an important part to play. Therefore, I'd spend money explaining what freedom, liberties, true holistic liberalism, openness, and democratic systems means. And tell the unfree people what the benefit is. I am working on such a project for Africa.

Education: Our brain can filter data and information for quick decision making. AI cannot real well, without a huge algorithm on top of its purpose of for example responding to a question. I`d spend that money on new forms of education for children, adults and like me old people. Topics like teaching critical thinking, ethics, law and methods in philosophy, complexity and systems theory must be thought to everybody.

Responsibility: Like Tristan Harris proposed, companies that create AI systems need to be accountable and responsible by law for potential damages their AI creates. See his film "the social Dilemma". Understand that this must be global and that's where the problem starts: China, Russia, the BRIC country would not support such an initiative of making and adhering to such a global law. However, the concept is clear: We need business laws that make AI producers accountable on a global basis for the risk and damage it could cause. Like any other business. Another law dearly deserves more usage: Anti trust laws to break the power of monopolies, especially platform companies.

Taxation: If AI replaces or will replace jobs, then it should be taxed for the work it does. Equivalent to the tax rates we taxpayers must pay. In certain countries up to 60%. And if those countries do not deploy such a tax system, you as an ethical based company, do not buy from them. This is a principal of Adam Smith's Theory of moral sentiments.

Stop Subsidies: https://www.theguardian.com/cities/2018/jul/02/us-cities-and-states-give-big-tech-93bn-in-subsidies-in-five-years-tax-breaks

Stop to Fix: Those few responsible people who cold should stop this race to fix and determine AI value add, ethic and socially responsible focus. Support those who want to repair and correct the legal, contractual, social, and conditional weaknesses of the internet,

like Tim Barner Lee with his Solid project. Hanna Arendt also proposed pauses to reflect and think.

Don't buy: China focuses on spying software and may be the western world should not buy at all. The force of the western world is us free people and our creative mind to come up with better solutions. Let us!

Sources of above opinions
Interview with Merefith Whittaker and Tonnie Grob. Schweizer Monat 2023
Katja Gentinetta in Juli 2023 "Der Pragmatismus". "Machen Sie doch einmal Pause."
https://melaniemitchell.me/
https://www.inrupt.com/solid