Risks and ethics of AI

Human life has always been influenced by technology. Agriculture, the printing press, the steam engine, Internet and online payment are just some of the few examples influencing our life. Artificial Intelligence is just the next stage in the technology development. It enables change on an unprecedented scale and at a very high pace. Besides, it brings a new component into our reality—smart machines, capable of learning and acting without human input.

By Anastassia Lauterbach

DR ANASTASSIA LAUTERBACH
is Technology Enterpreneur, founder of 1AU-Ventures, Investor and Non-executive Director. She is Senior Advisor on Artificial Intelligence at McKinsey and Board Member at Dun & Bradstreet and Censhare. Dr Lauterbach trains boards in cybersecurity and cognitive / AI related technologies, and their links to corporate governance.

Unanticipated actions of AI agents and robots might lead to unprecedented legal claims, fear, and radical judgments on what the potential of these technologies might be.

MISTAKES IN DESIGN

As technology reflects its creators, it is not immune to biases due to lack of diversity in gender, age and ethnics, or lack of involvement of non-engineers to consider important aspects of usability, ethics, and ultimately, safety. For example, when translating from Turkish—a language that uses a single gender-neutral pronoun 'o' instead of 'he' or 'she'— Google Translate attributes 'he' to soldiers, doctors and entrepreneurs, and 'she' to nurses and teachers. Alexa struggles to understand different accents. FAIR Lab reports that their systems on matching people and professions tagged the picture of Obama with "basketball". A 2016 report from the Obama Office of Science and Technology Policy warned that the impact of ML powered algorithms on work has the potential to worsen inequality.

MALICIOUS INTENT

Some unscrupulous manufacturer might insert some unethical behaviors into their smart machines in order to exploit users for financial gain, or cover up for bigger issues of the technology (just think about the VW diesel scandal of 2015). Like in many traditional industries, reputational risk of such decisions might be enormous. Today new ethical standards are emerging, e.g. BS 8611 Guide to the Ethical Design and Application of Robots and Robotic Systems, or IEEE P700X "human standards" which would also support OEMs and ODMs in the ethical application of robots.

CYBER-HACKING

Adversarial attacks on data models, hacking into a fleet of self-driving cars to damage or even kill passengers, faking video images or voice are realistic scenarios, demanding new ways and solutions to minimize the cybersecurity risk.

MACHINES WITHOUT A HUMAN IN THE LOOP

AlphaGo Zero achieved superior skills at Chess, Shogi and Go, and outperformed the original AlphaGo program in December 2017 by completely removing the human input. For the first time, an artificial system demonstrated capabilities to learn on its own. An AI technology called 'metalearning' or AutoML grew in prominence, with systems being capable of optimizing the hyper parameters and then running learning algorithms within them. Besides, ML allows for systems to reuse data and experiences from other tasks when trying solving new tasks. This implies things can be created from scratch, from zero prior knowledge and human intervention. AI practitioners expect that in the future coding will be dominated by automated systems without significant human involvement.

Technology, being neutral, enables progress, but it might also cause harm. Planning and experimentations are much needed ingredients, if we want AI to help us, rather then take our jobs, degrade our creative skills, and deprive us from critical ability to question, argue and make independent decisions.

Since AGI and ASI are decades away, most people might think there is no need to worry about implications at this point in time. Nevertheless, even narrow AI brings challenges to life and societal order, as we know it. Countless publications quote job losses, a need to adopt a Universal Basic Income to cover necessities, and the benefits of offering lifelong educational opportunities to escape stagnation and large inequalities. According to Max Tegmark, if humanity wants to win the race for safe and beneficial AI, we need to fund AI safety and ethics research today.

Interdisciplinary collaboration and diversity might be the only answers to ensure we achieve safe and beneficial AI. Deep interdisciplinary collaboration will ultimately touch four major themes, representing the foundation of AI ethics: machine vs human goal alignment, decision-making, incentives, and safety.

Consideration of ethical questions in AI research and development is linked with practices of knowledge sharing and distribution. "Inequality in who gained from computers has been less about inequality in understanding key insights about computers, and more about lumpiness in cultures, competing standards, marketing, regulations," writes Robin Hanson on November 22, 2008. In this context, geo-political considerations and concentration of AI R&D in the hands of what I call "full-stack AI companies" Alphabet, Apple, Microsoft, Facebook, Amazon, Baidu, and Tencent can't be neglected.

The Next Generation AI Plan of July 2017 declares Chinese Communist Party ambition to transform China into a global leader in AI, while achieving a domestic AI market of $150 billion by 2030. At the same time, Trump's White House cuts subsidies for research and policy development in AI, pointing out that the topic is not of the highest priority for the new administration. Ironically, while US Internet giants start opening up on developing ethical standards, supporting research around AI safety, similar initiatives are not visible in China.

In the future, those countries and companies will lead in safe and beneficial AI that reach a healthy balance between business, government and citizens. "This requires networked thinking and the establishment of an information, innovation, product and service "ecosystem." In order to work well, it is not only important to create opportunities for participation, but also to support diversity. There is no way to determine the best goal function: should we optimize the gross national product per capita or sustainability? Power or peace? Happiness or life expectancy? Often enough, what would have been better is only known after the fact. By allowing the pursuit of various different goals, a pluralistic society is better able to cope with the range of unexpected challenges to come."

Questions on decentralized design of AI systems, research around "self-criticism" of AI models and algorithms, prevention of adversarial attacks, improvement of data quality to enable user-controlled information filters, emphasis on collaborative opportunities of diverse teams along with increased digital literacy of corporate boards, and general public will not be solved easily.