УДК: 537.8 (07) (043)
Луцишин Р. - ст. гр. СІ – 31
Тернопільський національний технічний університет імені Івана Пулюя

BENEFITS & RISKS OF ARTIFICIAL INTELLIGENCE

Supervisor: Perenchuk O.Z.

From SIRI to self-driving cars, artificial intelligence (AI) is progressing rapidly. While science fiction often portrays AI as robots with human-like characteristics, AI can encompass anything from Google's search algorithms to IBM's Watson to autonomous weapons.

Artificial intelligence today is properly known as narrow AI (or weak AI), in that it is designed to perform a narrow task (e.g. only facial recognition or only internet searches or only driving a car). However, the long-term goal of many researchers is to create general AI (AGI or strong AI). While narrow AI may outperform humans at whatever its specific task is, like playing chess or solving equations, AGI would outperform humans at nearly every cognitive task.

In the near term, the goal of keeping AI's impact on society beneficial motivates research in many areas, from economics and law to technical topics such as verification, validity, security and control. Whereas it may be little more than a minor nuisance if your laptop crashes or gets hacked, it becomes all the more important that an AI system does what you want it to do if it controls your car, your airplane, your pacemaker, your automated trading system or your power grid. Another short-term challenge is preventing a devastating arms race in lethal autonomous weapons.

In the long term, an important question is what will happen if the quest for strong AI succeeds and an AI system becomes better than humans at all cognitive tasks. As pointed out by I.J. Good in 1965, designing smarter AI systems is itself a cognitive task. Such a system could potentially undergo recursive self-improvement, triggering an intelligence explosion leaving human intellect far behind. By inventing revolutionary new technologies, such a superintelligence might help us eradicate war, disease, and poverty, and so the creation of strong AI might be the biggest event in human history. Some experts have expressed concern, though, that it might also be the last, unless we learn to align the goals of the AI with ours before it becomes superintelligent.

Most researchers agree that a superintelligent AI is unlikely to exhibit human emotions like love or hate, and that there is no reason to expect AI to become intentionally benevolent or malevolent. Instead, when considering how AI might become a risk, experts think two scenarios most likely:

The AI is programmed to do something devastating: Autonomous weapons are artificial intelligence systems that are programmed to kill. In the hands of the wrong person, these weapons could easily cause mass casualties. Moreover, an AI arms race could inadvertently lead to an AI war that also results in mass casualties. To avoid being thwarted by the enemy, these weapons would be designed to be extremely difficult to simply "turn off," so humans could plausibly lose control of such a situation. This risk is one that's present even with narrow AI, but grows as levels of AI intelligence and autonomy increase.

The AI is programmed to do something beneficial, but it develops a destructive method for achieving its goal: This can happen whenever we fail to fully align the AI's goals with ours, which is strikingly difficult. If you ask an obedient intelligent car to take you to the airport as fast as possible, it might get you there chased by helicopters and covered in vomit, doing not what you wanted but literally what you asked for. If a superintelligent system is tasked with a ambitious geoengineering project, it might wreak havoc with our ecosystem as a side effect, and view human attempts to stop it as a threat to be met.