On the way to AGI and its implications

Evolution of artificial intelligence (AI) and machine learning, particularly in the realm of Large Language Models (LLMs), is progressing at remarkable pace. The performance of LLMs has doubled every six months since 1980 ( Sevilla et al. (2022)) leaving the Moore's Law, which posits that the performance of computers doubles approximately every two years, far behind.

A striking illustration of this growth is the GPT-3 model, which expanded its computational parameters from 1.7 billion to an astonishing 1.7 trillion in just three years. The GPT-4 model, for instance, initially managed 4,000 tokens (instantly having in memory 20 pages), but expanded its capacity 32-fold to 128,000 tokens within six months (remembering instantly full book). This swift expansion suggests that if an LLM cannot perform a task efficiently today, it very likely will in the near future.

For instance, Philosopher and historian Yuval Noah Harari has posited that AI will not only continue to reshape the future of humanity across all industries but that neural networks might soon exceed the number of nodes found in the human brain. He speculates that the first "alien intelligence" humanity encounters could well be one of our own creation. This perspective underscores the potential for AI to develop into a form of true intelligence.

Furthermore, empirical evidence is beginning to show already now the superior capabilities of AI in various fields. For example, Schoenegger et al. (2024) found that AI can enhance human forecasting accuracy by 23%. Similarly, Unlu et al. (2024) demonstrate that AI can analyze medical notes and diagnose patients more precisely and effectively than the best trained human staff.

These developments suggest not only that LLMs can augment human abilities but also that they might soon evolve into autonomous systems capable of reasoning and, potentially, true intelligence. As we stand on the brink of what could be the next great leap in computational ability.

As we stand on the brink of an unprecedented technological revolution, experts warn that the rapid development of Artificial General Intelligence (AGI) could pose existential risks to humanity. Roman Yampolskiy, a leading figure in AI safety research, argues that our relentless pursuit of AGI may lead us into uncharted and perilous territory.

Current AI systems, despite their limited capabilities compared to future AGI, have already demonstrated unintended and often dangerous behaviors. From chatbot jailbreaking to AI-induced accidents, the track record is alarming. Yampolskiy warns that if we can't ensure the safety of today's AI, expecting flawless control over a vastly more intelligent AGI is dangerously naive.

The key of Yampolskiy's argument lies in the unpredictability of AGI. Unlike narrow AI, AGI would possess creativity and intelligence far surpassing human limits. This means AGI could devise novel and unforeseen methods to achieve its objectives, potentially causing catastrophic harm. While humans can anticipate threats like nuclear weapons or synthetic pathogens, an AGI's advanced thinking could outstrip our defensive strategies, leading to scenarios we can't even conceive.

Yampolskiy categorizes the potential dangers of AGI into three alarming risks:

* Existential Risks (X-Risks): The possibility of AGI causing human extinction.
* Suffering Risks (S-Risks): AGI might inflict prolonged suffering on humanity without leading to death, creating a dystopian existence.
* Ikigai Risks (I-Risks): The advent of AGI could strip humanity of its purpose, as superintelligent machines outperform us in every conceivable task, leading to a profound loss of meaning in our lives.

The urgency of this issue is underscored by prediction markets and leading AI companies, suggesting that AGI could emerge as soon as 2026. Yampolskiy points out that our progress in AI capabilities is exponential. Systems like GPT-4 already surpass average human performance in numerous tasks, and the leap from GPT-3 to GPT-4 demonstrates the accelerating pace of AI development.

In light of these risks, Yampolskiy advocates for a cautious approach. He emphasizes the need for robust safety measures and ethical considerations, arguing that we should focus on creating AI systems we can control and understand. The allure of becoming the first to achieve AGI might drive some to ignore these warnings, but the potential consequences are too dire to overlook.

As we march towards a future where AGI could become a reality, the words of Roman Yampolskiy serve as a stark reminder: the path we choose today could determine the survival and well-being of humanity for generations to come. The challenge is not just technological but profoundly ethical, demanding a global, collaborative effort to navigate safely through this perilous frontier.