Could we create a morally good AI, which would never threaten us?

Tony Czarnecki, Sustensis

London 29/11/2023

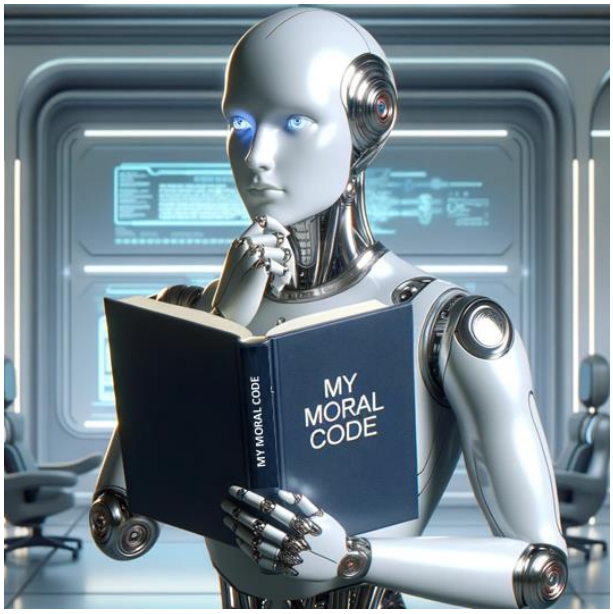

Image credit: DALL-E

To mitigate the risk of Superintelligence, which I perceive as a single, global, most advanced AI system, acting against our interests or even becoming outright malevolent, we must exercise early control over its development as it becomes increasingly more intelligent. To do that, we need a global and immediate implementation of the mechanisms controlling Superintelligence. This requires diverse approaches, which may collectively, better control the evolving "mind" of Superintelligence.

One such innovative approach is proposed by Yann LeCun's, Chief AI scientist at Meta. His views on controlling AI are optimistic, including solving the so called alignment problem, i.e., aligning AI's goals and motives with human values and preferences. He maintains this opinion in an interview with Financial Times made in October 2023, where he suggests that "several 'conceptual breakthroughs' were still needed before AI systems approached human-level intelligence. But even then, they could be controlled by encoding 'moral character' into these systems in the same way as people enact laws to govern human behaviour." This is broadly in line with the opinion of another super optimistic AI scientist, Gary Marcus. It contrasts with the prevailing view among AI researchers who maintain that controlling a superintelligent AI might be impossible, as it is impossible for a monkey trying to control a human. That was one of the reasons why on November 1st the Global AI Safety Summit was convened at Bletchley Park in the UK.

However, LeCun's proposal focusing on encoding a "moral character" into AI systems, ensuring they act ethically towards humans, deserves a closer examination. This idea is based on the possibility that AI's intelligence and its goals can be decoupled, allowing the development of AI systems that are intelligent but driven primarily by goals aligned with human values. While this concept sounds theoretically feasible, implementing it in practice remains a significant challenge. However, irrespective of the feasibility of the method he proposes, it is an interesting and potentially valuable approach to controlling AI.

In an article '**Want to Ensure AI Never Threatens Humanity? Make It Be Good**' discussing that interview, Alberto Romero raises two caveats to LeCun's proposal. First, relying on external control mechanisms, like laws, might not be effective for superintelligent AI. Instead, moral principles should

be fundamentally encoded into the AI's design. Secondly, the concept of morality is subjective and varies among humans, making it difficult to create a universal moral character for AI.

On the other hand, implementing morality as a parallel backbone to the advanced AI decision-making may be easier than creating a superintelligent humanoid in the context of Moravec's Paradox. In his book published in 1988 **'Mind Children: The Future of Robot and Human Intelligence'** Moravec postulates that it is easy to train computers to do things that humans find hard, like mathematics and logic, but it is hard to train them to do things humans find easy, like walking and image recognition. Morality does indeed fall into the first category, since like higher cognitive functions, it is a relatively recent evolutionary development and might be easier to replicate in AI than more ancient, optimized human skills. Although I share LeCun's optimism, like Alberto Romero, I also think that the practicality of implementing such a system remains uncertain and doubtful.

The first problem, linked with practicality, lies with agreeing human values, the cornerstone of morality. Considering the current global politics this boils down to the following questions: what type of morality can be considered as human-generic, who would define it, and how long would it take to agree the common human morality. A short answer is – it is unrealistic to expect it could be ever done. If it were at all possible that all states agree on something so fundamental to their identity, it would take decades to achieve that, whilst the new algorithms for humans' morality would need to be developed in a few years' time. There may be a slight possibility to agree and develop 'a narrow morality' algorithm broadly acceptable by many countries but not by all, once we have a de facto World Government, rather than a truly global government.

Secondly, morality may have not developed until consciousness has reached a certain level. That is why it is only present in humans, and perhaps to some extent, in apes or octopuses, the subject not raised neither by Yann Le Cun, nor Alberto Romero. Overall, it is an innovative proposal that should be implemented with all other methods, such as those proposed by Nick Bostrom in his seminal book 'Superintelligence'. However, none of them guarantees a failsafe control. We can only increase the probability of effective control by applying all feasible methods together. But even if we could create a morally good AI reflecting best traits of a human character, at some stage a superintelligent being may see the human life purpose and the implementation of those morals very inconsistent. For example, if we set those morals based on the UK law in the 1950' death penalty was perceived as absolutely moral but since 1969 it has been deemed immoral. Human morality is culture dependent and has been changing over the millennia. The AI's morality would not be stable either. To propose a moral code for AI and hoping it will save us from AI becoming malicious is at best a wishful thinking.

Therefore, embedding a moral code int AI may only be a stop gap. It would be better than nothing but similarly, as all other methods of controlling AI, it is based on hope that humans will be able to control AI indefinitely. My view is that it is a forgone conclusion that sooner or later we would be the losers in this struggle for dominating the world. Instead, we should accept that AI is the next step in human evolution. The biological homo sapiens will be gone. However, we may be the first ever creation of nature, which has designed its own evolution into a new species – a digital homo sapiens. If we accept that notion, then a logical approach is to start a civilisational transition to coexistence between humans and AI in a tightly coupled physical metamorphosis, similar to a caterpillar becoming a butterfly. That approach has been advocated by Sustensis – a Think Tank on Civilisational Transition the Coexistence with Superintelligence. Let me briefly explain the concept.

The core of my proposal is to create, what I call, the **Master Plate**, a method which may be more effective, unless physics and biology make its implementation impossible. The Master Plate is based on a BCI-fused control of Superintelligence, the most advanced AI, by Transhuman Governors. They would be carefully selected (including socio-psychological profiling) and connected in a ring via exponentially improving BCI devices to hundreds or even thousands of other licensed Transhuman Governors. One element of that ring would be the Master Plate's 'control hub'. This is a hardware/software device similar to a computer's BIOS (Basic Input Output System) enabling

Transhuman Governors to control with their thoughts the main goals or decisions to be made by Superintelligence.

Such an approach would solve most the problems related to creating emotional, conscious and superintelligent being. But it would also start a gradual transformation of humans initially into Transhumans and ultimately into Posthumans – an entirely digital species. There would be no need for controlling AI, because it would be part of the most advanced Transhumans, as they would be part of the maturing Superintelligence.

The principle of AI's emotion, intelligence, and morality advancing in parallel with ours, where we are more advanced in consciousness and morality but less in intelligence, may be the best and the safest option because it would greatly reduce the problem of lack of global agreement on human values and morality. The only requirement would be to initially select the avantgarde of humans (Transhumans) by an independent body authorized by a de facto World Government.

I explain the detail of such a method of control, supported by several diagrams, in the context of a civilisational transformation, in my recent book: '**Prevail or Fail – a Civilisational Shift to the World of Transhumans**'. But those who are only interested in the control aspects of AI may read my article - '**The Master Plate – Controlling AI from within**'.