A Thought Experiment

Thomas Walker

December 2022

Suppose you find yourself stuck in the stone age. With nothing but the intellect, you have acquired as being part of a lineage of descendants where multiple conceptual leaps in thinking have occurred and the minimal garments providing protection to the elements. As a result of many years of evolution, your brain possesses the capabilities that sees the construction of skyscrapers and the operation of large-scale governance. However, that is all left behind in your old life, you are now living in the stone age. So, what do you do? Your first instinct is to reach into your pocket... but then you realise your garments have no pockets. How can that be possible you think, where else would your phone be? You are left in the stone age with only the knowledge of such technologies, but you do not have access to any of it. You have now realised you are stripped of many of the amenities that you have become accustomed to. So, what do you? It is probably important to establish some essentials such as food, water, and warmth. With these goals in mind, you collect some firewood and locate yourself by a stream with a hand full of berries you scavenged from some nearby bushes. You now sit around your lit campfire; cook the rabbit you captured earlier and contemplate your next steps. Your priority is to survive and gain access to resources to let you reconstruct your previous life. Frantically, you jot down your plans including details of technologies with the fear that you may forget them before you are able to recreate them. It is getting late, so you go to sleep somewhat comforted by the fact that you have a developed plan, but apprehensive for the situation you currently find yourself in.

From afar a group of stone age hunters are observing your behaviour. You seem very familiar to them, although a bit quirky. They notice you prioritize objectives such as staying warm and having access to water. You also have been able to harness fire to provide you with warmth, although they mocked you as it took you a while to get it going. However, they were puzzled by the fact that you were burning your catch of the day on this fire. They had never considered this act before, but they were not able to comprehend its importance and simply regarded it as one of your quirks. They continued to look on as you slept, the act of you sleeping further convinced them that you may be one of them and a friendly creature to approach.

Suddenly, a gust of wind whipped through the fire, sending sparks crackling high into the air. The wind caught, what seemed to the hunters, a thin piece of bark. They cast their eyes down on the bark and noticed a string of odd shapes. The induvial symbols seemed chaotic, but they were neatly ordered on the piece of bark. The hunters didn't have the capacity to understand what you had transcribed on that piece of bark. The discovery of this only morphed the hunter's confusion about you into fear. Was this creature really one of us? Does it experience pain, thirst, or hunger? Can we trust it?

I write this story to draw an analogy with the field of research that aims to develop artificial intelligence systems that are smarter than humans. Many are working on this challenge of creating artificial general intelligence (AGI) systems that have the capacity to excel at many of the tasks humans can do as well outperform humans on many other tasks. Such a system will inevitably have many consequences, and this is the main topic of research within AI safety which I am curious about. The human in this story corresponds to the AI system, and we humans are represented by the stone age hunters. The fact that the human has many goals aligned with the hunters (i.e., staying warm and having a water supply) is equivalent to the instrumental goal hypothesis in AI research. Which states that such a generally intelligent system will have to pursue a set of instrumental goals, such as collecting resources, to achieve its main objectives. The cooking of the rabbit done by the human is analogous to an AI system capitalising on current technologies in a way that we cannot imagine due to our limited intellect. Furthermore, such a system will operate on a different conceptual level to us so will produce outputs that we cannot comprehend, represented in the story by the notes jotted down on the piece of paper. Then finally, the fear experienced by the hunters is akin to the fear

that many AI researchers that such a power system may have goals that are not aligned with humans' goals which could lead to such a system posing a threat to humanity.
This field of research begins to ask questions as to whether we can trust such an intelligent system to have goals that are aligned with our own. If we were to let it operate in our society, would it work with us, or suppress us to achieve its own goals? AI safety research aims to find ways to ensure that AGI systems have goals aligned with our own so that we can utilise this technology for the many benefits it provides without having concerns about its potential harm.