# Research Proposal

Michaël Trazzi

February 25, 2019

## The Treacherous Turn

At some point in the development of a seed AI, it may realise that it needs to get rid of its supervisors to achieve its goals. To maximise its chances of taking over, the seed AI might initially exhibit behaviours desirable and non-threatening to human designers, before undertaking a *treacherous turn* when humans are no longer a threat[1]. From the human perspective, the AI would continue to exhibit desirable behaviours, and by the point that humans recognize it to be dangerous, it might already be unstoppable.

The treacherous turn raises challenges at the intersection of AI Boxing and Interpretability. How can humans limit an AI's abilities to deceive its supervisor[2][3]? How can we develop interpretable learning methods that could prevent an AI from pretending to be subhuman in intelligence, while it concurrently manages to become superintelligent? This might, for instance occur, if the AI obfuscates its code after a first moment of vulnerability[1].

My research will build on my previous simulation of the treacherous turn[4]. More precisely, I will generate multiple gridworlds with the same core structure (similar to my previous simulation) to test whether an agent learning with meta-reinforcement learning (meta-RL) as described in Wang et al.[5] could learn to commit a treacherous turn without ever trying to kill its supervisor. In my existing research I have already reproduced some results of the mentioned meta-RL paper[6], and I am in contact with the authors for further work using their algorithm.

Then, I will try to develop interpretability methods to empirically compare whether Goertzel's sordid stumble or Bostrom's treacherous turn seems most likely[7]. I could do this by visualising and analysing the meta-RL process. That would allow me to estimate the number of episodes necessary to meta-learn the concepts of *supervisor* and *deception*, and compare them to the number of episodes necessary to undertake a treacherous turn. My main concern with this approach is the instability of meta-RL methods and the tractability of running such algorithms on gridworlds. However, this is a valuable project because the treacherous turn is a problem that humanity will likely face until the end of the Human Era (i.e. period where Homo Sapiens is the dominant lifeform on Earth), so early research on the treacherous turn could become the foundation

for a long-lasting research agenda. Furthemore, the developed framework could be used for other Interpretability tasks.

## Instrumental Behaviour of an Agent Building its Model of Reality

Let's suppose an agent is planning to build its most accurate model of reality. An option would be to collect as much data as possible. Another option would be to infer new knowledge from running simulations (here, simulation is defined as "running an abstract model of the world using the laws of physics as input, without optimizing for inference architecture").

To what extent does gathering additional information have, on average, higher expected utility than running computer simulations (where the agent's utility function is defined in terms of value of information)? More precisely, how much time should an agent spend collecting new data (for instance by improving the precision of sensors, inventing new sensors or moving the position of already existing sensors), as opposed to inferring new knowledge from simulations?

A quantitative answer to those questions would be valuable because creating an accurate model of reality is likely to be a convergent instrumental goal[8]. Humans would henceforth be able to think ahead by making more accurate predictions about smarter-than-human agents' instrumental behaviours. For instance, if running computer simulations has (on average) higher expected utility, then we would expect computing power or better simulators to be a priority for smarter-than-human agents, as opposed to data collection.

In my research, I will start by characterizing the contexts in which running simulations would be preferred to data collection, and vice versa (e.g. in the case of hardware (resp. algorithmic) recalcitrance[1]?). In practice, I will try to get feedback by writing weekly blogposts on LessWrong and by bouncing some ideas off Anders Sandberg and Stuart Armstrong who have published related papers[9][10].

My main concern with this approach it the difficulty in estimating what portion of physics can be directly observed without ever moving around in the universe. For instance, it seems impossible to discover the existence of blackholes by just looking at a glass of water. However, I think there exist low-hanging fruits in precisely estimating the value of information per FLOP an agent could gain from running one simulation. Indeed, Shulman and Bostrom estimate the amount of computation necessary to recapitulate the clumsy evolutionary strategies that lead to civilization to be in the $10^{31}$-$10^{44}$ FLOPS range[11], which is a very large interval.

## The Limits of Brain-Computer Interfaces, and their Consequences

When it comes to the distant future of Homo Sapiens, the cognitive limits of enhanced humans are not well understood. Several paths could lead to an enhancement of human intelligence. For instance, Neuralink is developing brain-computer interfaces (BCI) for cognitive enhancement, and the path to whole brain emulations is becoming a real prospect[12][13][14]. However, there are walls of cognitive excellence that could be reached beyond which it is not possible for humans to transcend: humans might only upgrade their intelligence up to a certain intelligence ceiling. Thereafter, the only way to upgrade their intelligence would be to completely restructure it.

Studying the practical and theoretical implications of such a structural cognitive upper bound is valuable because it would shape future decisions, especially cause prioritisation (e.g. if brain-computer interfaces can only boost human intelligence up to a certain threshold, humanity might need to prioritize AI Safety, rather than invest in BCI technology).

In practice, I will start by finding all the projects related to brain-computer interfaces that address cognitive upper bounds. Then, for each of these projects (for instance, whole brain emulations), I will try to precisely estimate those cognitive upper bounds, in terms of both *qualitative* and *speed* intelligence[1]. To that end, I plan to solicit the feedback of neuroscience researchers that I have already contacted at DeepMind (such as Jane Wang and Matthew Botvinick), and work closely with Anders Sandberg. My main concern with this approach is that most of the state-of-the-art in BCI is currently mostly engineering, with little theoretical foundations (or at least those foundations are unpublished). Yet, even if we don't have the full picture on the inner workings of brain-computer interfaces, it's still possible to give approximate bounds for their limits. Indeed, for my previous work[3], I estimated the cognitive limits of human intelligence, sometimes reasoning from first principles.

## References


[1] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. 1st. New York, NY, USA: Oxford University Press, Inc., 2014.

[2] Stuart Armstrong. "Good and safe uses of AI Oracles" (2017). URL: `http://arxiv.org/abs/1711.05541`.

[3] Michaël Trazzi and Roman V. Yampolskiy. "Building Safer AGI by introducing Artificial Stupidity" (2018). URL: `http://arxiv.org/abs/1808.03644`.

[4] Michaël Trazzi. *A Link To The Past Gridworld Environment for the Treacherous Turn*. `https://github.com/mtrazzi/gym-alttp-gridworld`. July 2018.

[5] Jane X. Wang et al. "Learning to reinforcement learn" (2016). URL: `https://arxiv.org/abs/1611.05763`.

[6] Michaël Trazzi. *The two-step task*. `https://github.com/mtrazzi/two-step-task`. Dec 2018.

[7] Seth Baum, Anthony Barrett, and Roman V Yampolskiy. "Modeling and Interpreting Expert Disagreement About Artificial Superintelligence" (2018).

[8] Stephen M Omohundro. "The basic AI drives". 2008.

[9] Anders Sandberg, Stuart Armstrong, and Milan M Cirkovic. "That is not dead which can eternal lie: the aestivation hypothesis for resolving Fermi's paradox". *arXiv preprint arXiv:1705.03394* (2017).

[10] Anders Sandberg. *Space races: Settling the universe Fast*. 2018.

[11] Carl Shulman and Nick Bostrom. "How hard is artificial intelligence? Evolutionary arguments and selection effects". *Journal of Consciousness Studies* 19.7-8 (2012), pp. 103–130.

[12] Tim Urban. *Neuralink and the Brain's Magical Future*. 2017. URL: `https://waitbutwhy.com/2017/04/neuralink.html`.

[13] Anders Sandberg. "Feasibility of whole brain emulation". *Philosophy and Theory of Artificial Intelligence*. Springer, 2013, pp. 251–264.

[14] Anders Sandberg and Nick Bostrom. "Whole brain emulation" (2008).