<https://www.alignmentforum.org/posts/aqhMLqaoHb7uob7fr/if-i-were-a-well-intentioned-ai-iv-mesa-optimising>

Mesa-optimization > Mesa-optimizer can be controlled: “aligned mesa-optimisers are generally better than controlled mesa-optimisers, *but it is hard to tell the difference between an aligned and a dangerous unaligned mesa-optimiser*”

Mesa-optimization > Mesa-optimization can not be smarter than base: if it is not as smart, then it may act according to Ubase as opposed to the assigned Umesa, which then causes potential conflict between the mesa optimizer and the base optimizer.

Mesa-optimization > Mesa-optimizer can be corrigible: If a mesa-optimizer is corrigible, it can be controlled.

<https://www.alignmentforum.org/posts/rt5X74Az3mXwTubRA/trace-goals-and-principles>

No nodes, but may be useful for causal reasoning ideas, if we want to delve into that section, down the line

<https://www.alignmentforum.org/posts/ivpKSjM4D6FbqF4pZ/cortes-pizarro-and-afonso-as-precedents-for-takeover>

Likens AI to Cortes/Pizarro/Afonso – their minority but armed with better technology allows for total domination. More than this, “strategic and diplomatic cunning” lead to victory, which AI could foreseeably achieve. Furthermore, the Spanish conquistadors were in their own tumult and backstabbing landscapes (which is similar to AI now with large corporations)

<http://www.incompleteideas.net/IncIdeas/BitterLesson.html>

AI research clearly points to methods that leverage computation are most effective. Human’s tricks of representation short-term improve results, but long term do not. This was seen in chess, go, speech recognition, and computer vision.

<https://ai-alignment.com/iterated-distillation-and-amplification-157debfd1616>

A “scheme for training machine learning systems that can be robustly aligned to complex and fuzzy values,” based off Alpha Go Zero.

<https://www.alignmentforum.org/posts/3kwR2dufdJyJamHQq/mechanistic-transparency-for-machine-learning>

ML Scales to AGI > Mechanistic Transparency can be achieved: Mechanistic Transparency is a tool that takes a NN and outputs some sort of representation of what it is doing (such as pseudocode) that can be scalable (so it is a very high level representation, and then cascades down to mere nodes and edges) and not require the system to be run.

<https://www.alignmentforum.org/posts/JYdGCrD55FhS4iHvY/robustness-to-fundamental-uncertainty-in-agi-alignment-1>

“The distribution of risks from AGI suggest we can reduce the risk of catastrophe in attempts to build aligned AGI by trading off false positives for false negatives.

<https://arxiv.org/pdf/2002.06177.pdf>

ML Scales to AGI > Deep Learning’s abstractions can link to reality in a generalizable fashion: Can AI create abstractions that can apply to other scenarios in a more generalizable manner than the current state with “stark limitations, being brittle, data-hungry, unable to make sense of situations that deviate slightly from their training data or the assumptions of their creators, and unable too repurpose themselves to deal with novel talks without significant involvement from human researchers.”

Deep Insights Needed > Human safety can be modeled: If human safety can be modeled, then AI would be safe, but the question remains of how can it be modeled?