The “*AI alignment problem*” problem

Brian M. Delaney
Mind First Foundation

Expert views on controlling AI

Omohundro, *The Nature of Self-Improving Artificial Intelligence* (2007/2008)

How can we ensure that this technology acts in alignment with our values?

Boström, *Superintelligence* (2014)

In principle, we could build a kind of superintelligence that would protect human values.

Russell, *Human Compatible* (2019)

The machine's only objective is to maximize the realization of human preferences.

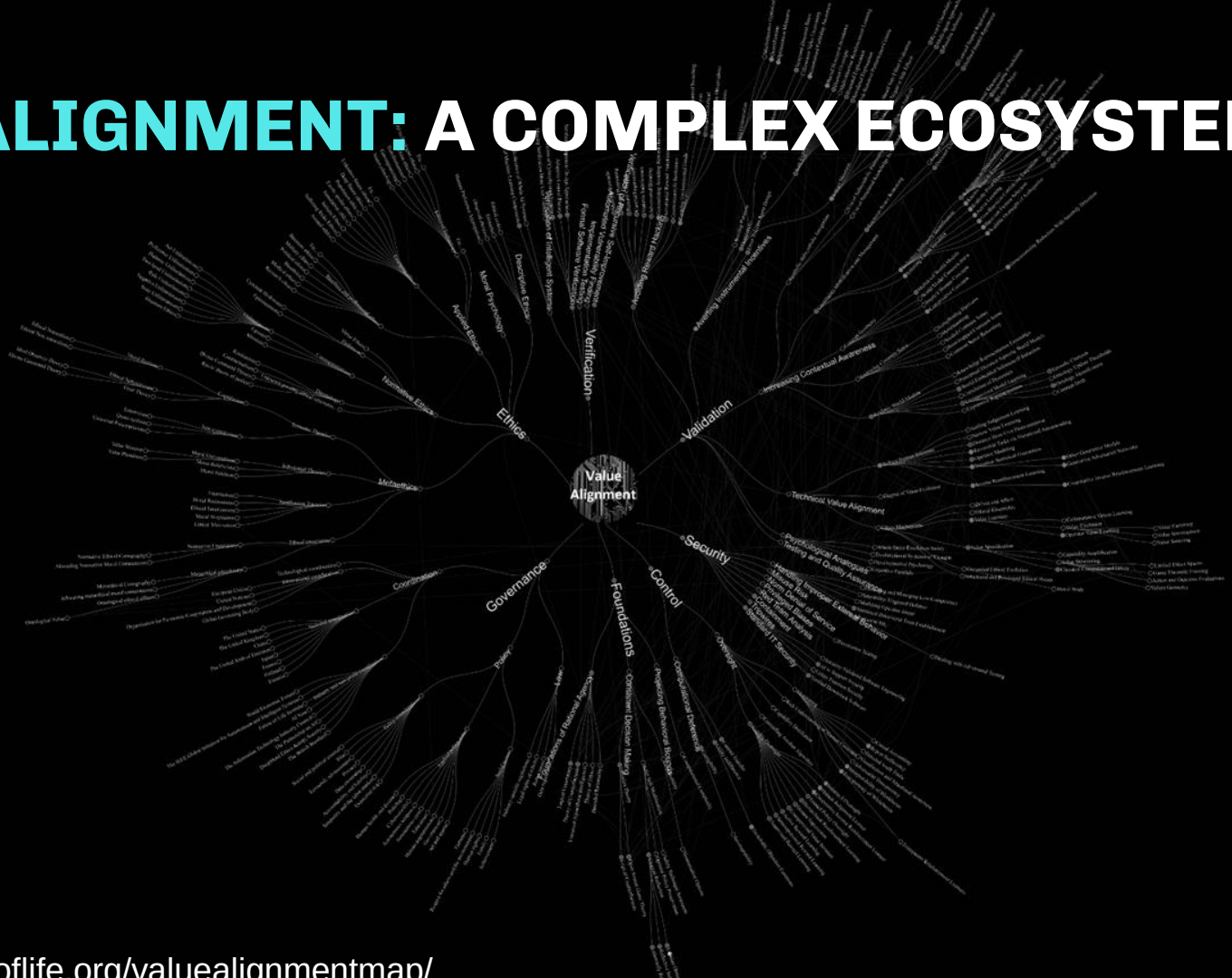
Russell *(ibid)*

I focus in particular on the problem of control: retaining absolute power over machines that are more powerful than us.

The problem of achieving agreement between our true preferences and the objective we put into the machine is called the value alignment problem: the values or objectives put into the machine must be aligned with those of the human.

— Russell & Norvig (2021). Artificial intelligence: A modern approach (4th ed.). p. 5.

ALIGNMENT: A COMPLEX ECOSYSTEM



https://futureoflife.org/valuealignmentmap/

alinen (15th cent.):

Copulate (of dogs, wolves).

— *https://www.etymonline.com*

Alignment to whom – *which values?*

COMMON REASONING ABOUT VALUES

1.
Human values vary

2.
We need to choose the “best of these human values”

3.
But how?

COMMON REASONING ABOUT VALUES

1.
Human values vary

2.
We need to choose the “best of these human values”

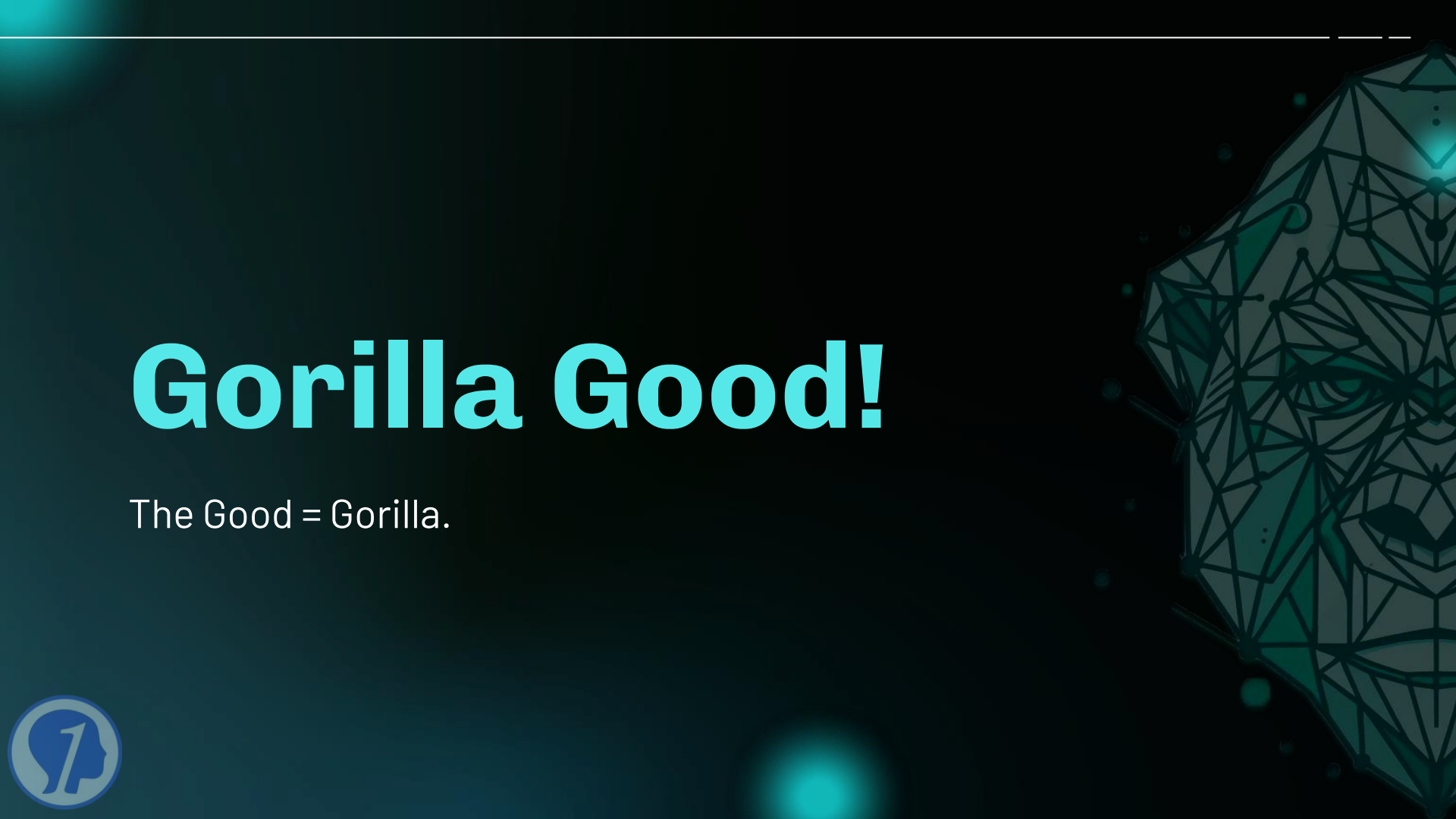
3.
To choose the best values, we need criteria that transcend particular humans

Around ten million years ago, the ancestors of the modern gorilla created [...] the genetic lineage leading to modern humans. How do the gorillas feel about this? Clearly, if they were able to tell us about their species' current situation vis-à-vis humans, the consensus opinion would be very negative indeed. [...] We do not want to be in a similar situation vis-à-vis superintelligent machines.

— Russell, *Human Compatible.*

Gorilla Good!

The Good = Gorilla.



“This is ridiculous!” “Quantum mechanics, microtonal music, Da Vinci!!” etc. “We are clearly The Good, and deserve to have superseded our banana-brained ancestors!”

“Human good!”

The Good = Human

Unsolved Ancient Problem:

How should we live our lives?

“Sum ergo cogito.”

—Nietzsche, *The Gay Science*

AI as Übermensch?

Thank you

Questions or comments?
brian@mindfirst.foundation

https://mindfirst.foundation

Acknowledgements

Mind First Foundation

Ranjan Ahuja
Preston Estep, PhD
Alex Hoekstra

OpenAI / DALL-E 3.0
Steve Omohundro
Stuart Russell
George Church
Dan Elton
Vitalik Buterin
Ted Bakewell
Microsoft

brian@mindfirst.foundation
https://mindfirst.foundation/presentations