Vague thoughts on AI alignment.

There is a fundamental analogy between AI and economies. The objective functions are consumer preferences, and the learning method (market mechanism) develops mesa-optimizers (firms) to achieve them.

The government has three jobs:

(1) Break monopolies

(2) Manufacture confidence

(3) Define and protect suitable property rights structures so that the “correct” preferences are represented.

This is kind of interesting in itself. The first two are about creating economic efficiency, but the third is a lot less “objective”. Disagreements over these aren’t really conflicts of ideology, but conflicts of interest. Governments that hyper-emphasize their own “political” preferences are considered fascist, governments that place some people as the property of others are considered slave states, etc.

The first two are not relevant to alignment, afaict.