The Tragedy of the AI Commons

Travis LaCroix & Aydin Mohseni

Background. Artificial intelligence promises to fundamentally change nearly every facet of our lives, for better or worse. In response to this reality, there has been a proliferation of policy and guideline proposals for ethical AI research. These documents are meant to specify 'best practices' to which engineers, developers, researchers, etc. ought to adhere. However, such documents are 'non-legislative policy instruments': they are meant to promote cooperation but are not legally binding. So, these reports are not intended to produce enforceable rules but are meant merely as guides for ethical practice. The proliferation of such guides raises pressing questions about their efficacy.

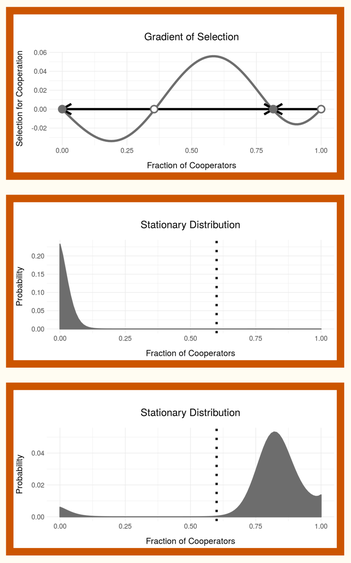

Fig. 1. Example qualitative dynamics for the gradient of selection (top) and stationary distribution (mid, bottom) of the mean-field dynamics

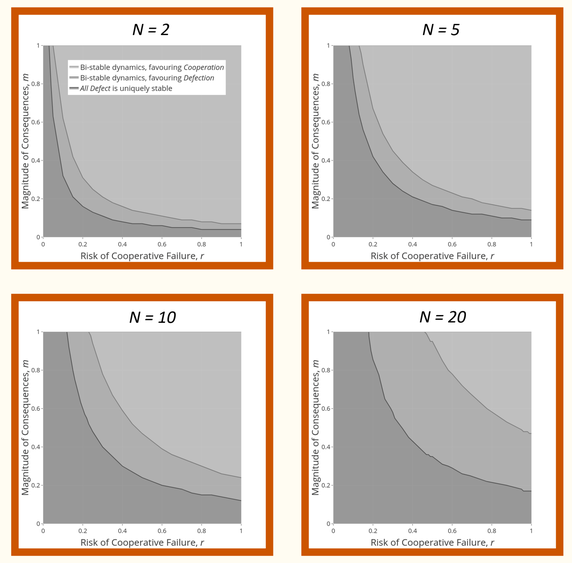

Fig. 2. Qualitative dynamics as a function of the risk, r, and magnitude, m, of the consequences of a failure to cooperate.

Game-theoretic analysis provides conditions for **cooperative success** of the **responsible development of artificial intelligence.**

Paper

Model

Moral 1. Lowering the cost of cooperation increases the likelihood of cooperative success.

Moral 2. Small, decentralised groups may benefit sustained cooperation for responsible AI research.

Moral 3. Voluntary participation in AI policy agreements may catalyse the spread of cooperation.

Moral 4. It is important to accurately figure both the risks and the consequences of non-cooperation.

Moral 5. Combining many proposals may undermine their prospects for success.

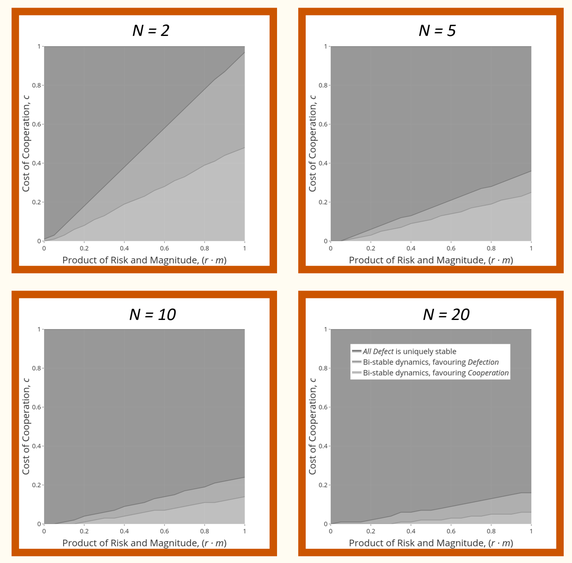

Fig. 3. Qualitative dynamics as a function of the product of the risk and magnitude of the consequences to fail to cooperate, $r \cdot m$, and the cost of cooperation, c.