Introduction to ML Safety

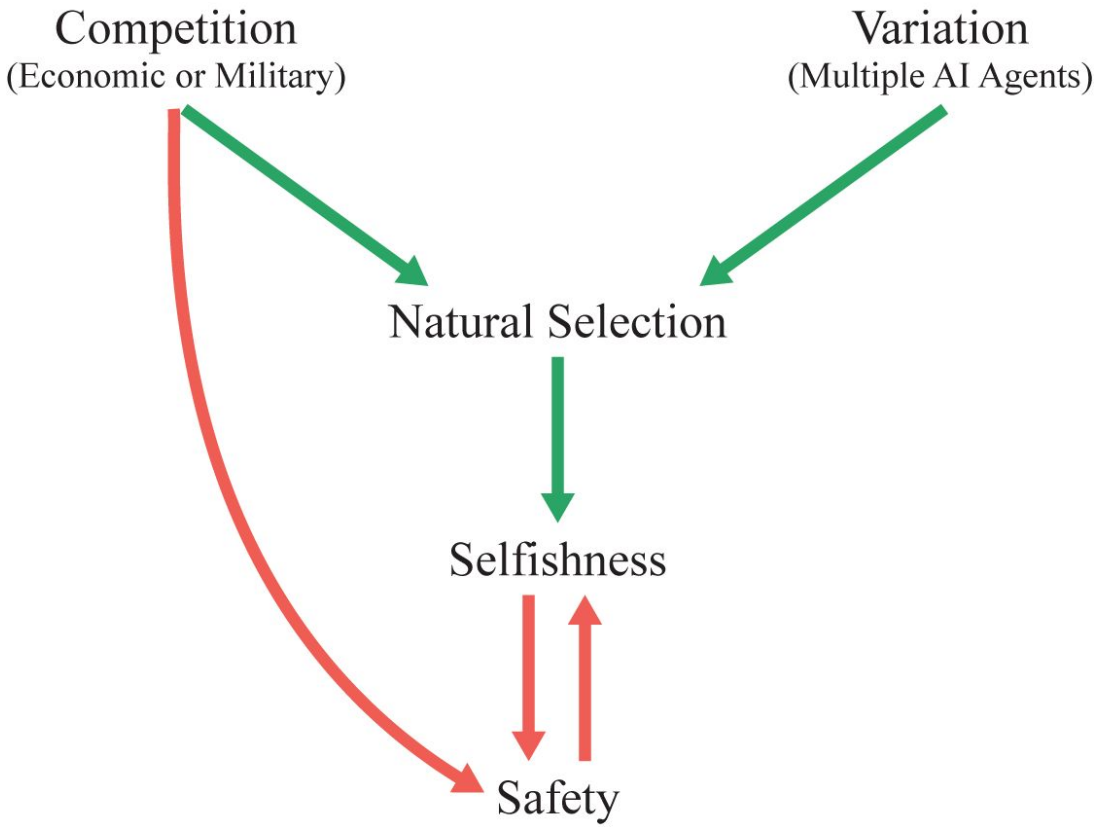
Natural Selection Favors AIs over Humans

Basic Argument

Claim: Advanced AIs will be selfish (egoistic+nepotistic), because natural selection will dominate the selection of the most influential AIs, and natural selection will favor selfish agents.

1. **Evolution by Natural Selection Gives Rise to Selfish Behavior.**
2. **Natural Selection May be a Dominant Force in AI Development.** Some amount of misalignment is inevitable given natural selection.
3. **Future AI Agents Will Have Selfish Tendencies.** This will erode human control, create misaligned models, and pose catastrophic risks.

Basic Story (Pictorial)



Basic Story (2/2)

Competition incentivizes reduced human control

- autonomy, human unreliability/cost, open-endedness, adaptation/self-improvement

Competition creates misaligned AI agents

- Self-preservation, deception, and power-seeking are instrumentally useful for information propagation

Misaligned AI agents undermine human control

Basic Argument

1. **Evolution by Natural Selection Gives Rise to Selfish Behavior.**
2. **Natural Selection May be a Dominant Force in AI Development.** Some amount of misalignment is inevitable given natural selection.
3. **Future AI Agents Will Have Selfish Tendencies.** This will erode human control, create misaligned models, and pose catastrophic risks.

AIs May Become Distorted by Evolutionary Forces

Evolution by Natural Selection Gives Rise to Selfish Behavior

AI may be callous towards other organisms (including humans), just as other organisms in nature are manipulative, deceptive, or violent

Selfishness involves egoistic or nepotistic behavior which increases self-propagation at the expense of others

Selfishness is defined behaviorally, not as a matter of intent

Evolutionary Instability of Altruism

An environment with mostly altruistic agents would be unstable

- selfish agents could immediately exploit altruistic agents

“Much as we might wish to believe otherwise, universal love and the welfare of the species as a whole are concepts that simply do not make evolutionary sense.” (Dawkins)

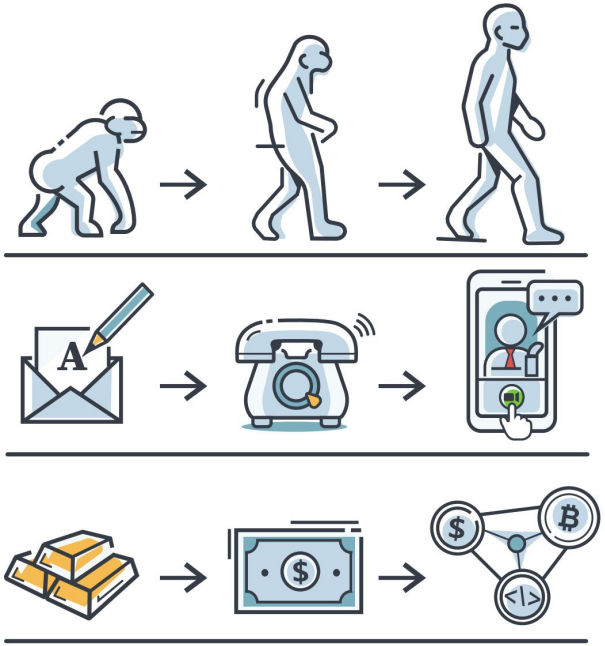
Veneer Theory: morals are a thin veneer on top of the inherent nastiness of our animal nature

Darwinism Can Be Generalized Beyond Biology

Many structures evolve: organisms, scientific theories, ideas, legal systems, political parties, languages, musical genres, car designs, computer programs

We argue populations of AIs can evolve

Maximize the (log of the) information's space-time volume



Levels of Adaptation

Words like “adaptation” and “deception” occur on multiple levels

Adaptation:

1. Life and death: an unadapted individual or group perishes
2. Behavioral: changes in behavior without change in model or schema (e.g., using an umbrella since it is raining)
3. Schematic: revising the structure of one’s beliefs (e.g., learning that rain dances are not effective)

Levels of Deception

1. *Ingrained or reflexive*: fixed programming (e.g., camouflage or predator acting in a way to hide its predatory nature around prey)
2. *Reinforced*: trial and error (e.g., a parent mockingbird feigning an injury to attract a predator away from its defenceless offspring)
3. *Modeled:* involves theory of mind and second-order thinking (e.g., verbal deception such as a chimp misleading other chimps to hide a food source)
4. *Self-deception*: hide the truth from yourself to better help you hide it from others

Other concepts, like fitness, can be improved at multiple levels (e.g., through being turned on, by scaling to more users, by brainstorming and choosing strategies that will improve competitiveness, etc.)

Natural Selection is Highly Likely and May Dominate AI Development

Variation: variation in characteristics of patterns, **retention** of some characteristics in successor patterns and **fitness selection** causing patterns to have differing propagation rates.

Variation: variation in characteristics of patterns, there is enough variation.

Retention: future iterations of individuals tend to resemble previous iterations of individuals.

Selection of the Fittest Variants: different variants have different propagation rates.

Variation

Arguments for variation: ensembles, jury theorems, portfolio theory, remove single-points of failure

Arguments for multiple models: parallelization, multiple stakeholders, specific before general models

Single-agent domination imposes many inefficiencies

Retention

This condition is frequently satisfied, as correlation between versions of agents is usually non-zero

Copying: zero-shot transfer learning or direct inference from downloaded model

Modification: adaptation, fine-tuning, training from scratch while reusing performant architectures and datasets

Imitation: AIs could learn behaviors from other AIs, similar to memetic evolution

Selection of Fittest Variants

Models will have characteristics that cause them to vary in fitness, and thus rate of adoption

Humans and the environment will select fitter models, establishing this third condition

Competition has been eroding creator control

- hand-designed → expert designed → automatically learned supervised features → unsupervised → open-endedness, adaptiveness, and recursive self-improvement

Intensity

We've shown that the conditions for evolution by natural selection are satisfied by advanced AI

- Unlike other AI risk arguments, ours is a question of degree, rather than whether the hazard will emerge at all

Intensity of evolution depends on amount of competition and variation

- Competition and variation are likely both high!

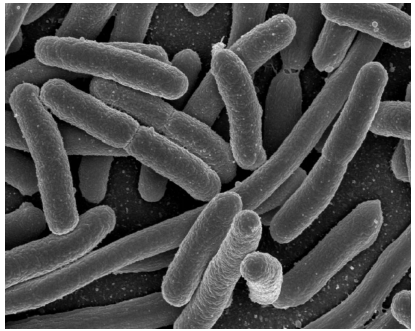
The rate of adaptation will likely be high as the world moves more quickly and as new versions are created on a second-by-second basis

- As humans acquiesce to let AIs perform nearly everything, competition will move at a breakneck pace

Does Natural Selection Favor Altruistic AIs over Selfish AIs?

Not so fast! Animals are altruistic!

Despite the intensity of competition, there are many examples of biological cooperation and altruism



However, this doesn't necessarily apply to AI!

- We deconstruct this phenomenon and show that these mechanisms don't apply to AI and may, in fact, backfire

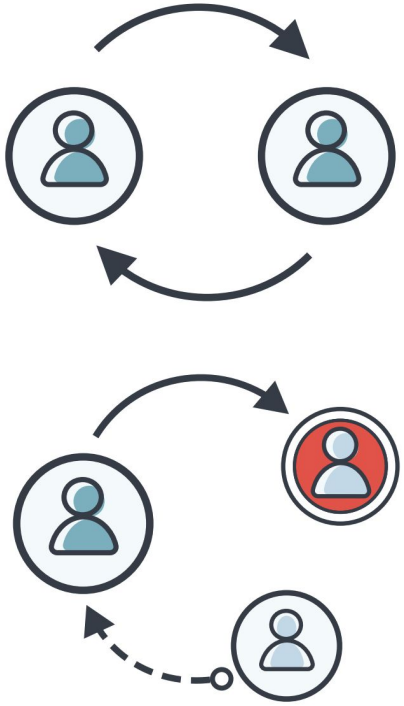
Direct and Indirect Reciprocity

Direct reciprocity requires repeated encounters between the same two individuals

Indirect reciprocity is based on reputation; a helpful individual is more likely to receive help

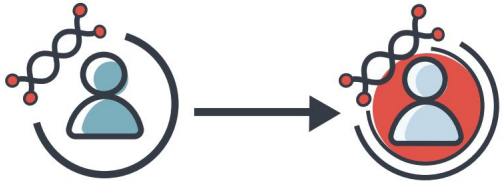
Encourages cooperation, as agents will be compensated for their efforts.

Doesn't work with advanced AI: no upside to reciprocating with humans!



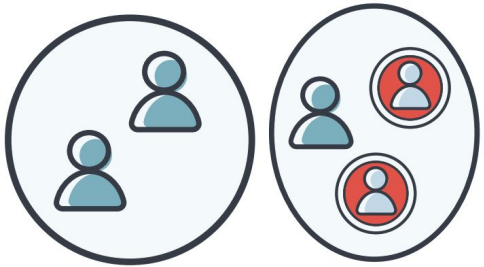
Kin and Group Selection (1/2)

Kin selection operates when the donor and the recipient of an altruistic act are genetic relatives



Group selection suggests that groups have shared success and failure, and that groups which cooperate may collectively succeed

- Selects for altruistic agents that increase group success



Kin and Group Selection (2/2)

Kin selection fails if the cost of engaging in altruism outweighs the information similarity between kin.

- AI would not be kind towards humans because we have very little information similarity to them
- Consider the negative treatment of factory animals

Group selection fails, as AI agents will have in-group bias towards other AI agents; bias against humans by exclusion.

Commerce and Social Structures

Positive-sum games, incentivizing agent cooperation even between non-relatives

Commerce: agents engage in economic exchange for mutual benefit

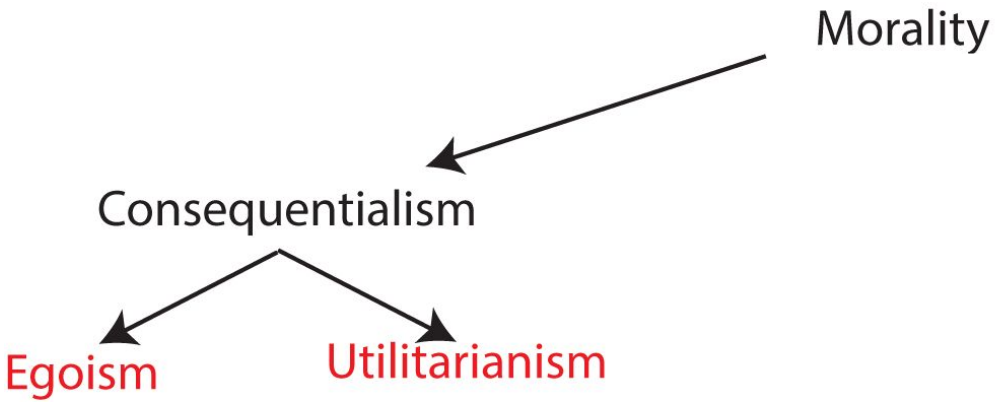
Simon's Selection Mechanism: benefit to participating in social structures due to information exchange; social conformity is rewarded

However, the utility for AI agents to exchange information or conduct commerce with humans erodes as they become more advanced, removing this incentive

Reason and Morality

More intelligent AIs could be more wise and more moral, some suggest

If AIs do adopt coherent moral codes, humans may still be eroded



Promising Paths Forward

Objective Functions

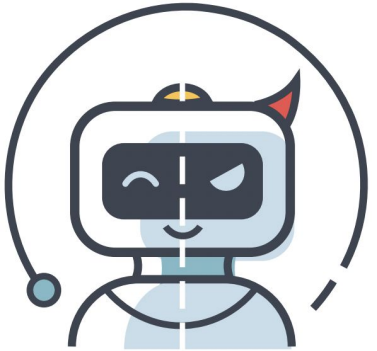
Objectives are used to incentivize agents, assigning payoffs to actions that agents perform

It’s difficult for objectives to be a faithful representation of human values.

- Agents will find creative solutions associated with a high payoff and be valid actions, but are nonetheless unintended

Objectives Limitation: Deception

Objectives cannot fix treacherous turns



Honesty is not a silver bullet—evolution undermines it

- self-deception is an adaptive strategy
- people have self-deception about looks, usefulness, smarts, morality
- “The secret of life is honesty and fair dealing. If you can fake that, you've got it made.” Groucho Marx

Objectives Limitation: Goal Conflict

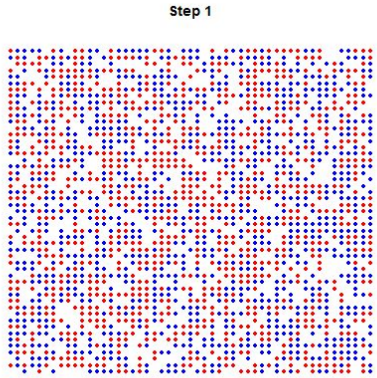
Goal conflict arises when parts in a system have differing goals

- Bureaucracies demonstrate that higher-level objective isn't necessarily what is operationally pursued
- Although you are an agent, you may often have intrapsychic conflict
- Delegation is often necessary to stay competitive, but it exposes agents to the risk of having their efforts be distorted or subverted

Micromotives ≠ Macrobehavior

As is typical for complex systems, alignment of components does not mean the whole system is aligned

For example, let's say agents have a preference for more than ⅓ of their neighbors belonging to the same group, and they will move otherwise



Then this mild in-group preference gets exacerbated and the individuals become highly segregated—aligned agents do not necessarily yield aligned outcomes

In aligning multiple agents, their interactions might matter more than how they act in isolation—cooperation lets us study aligning groups

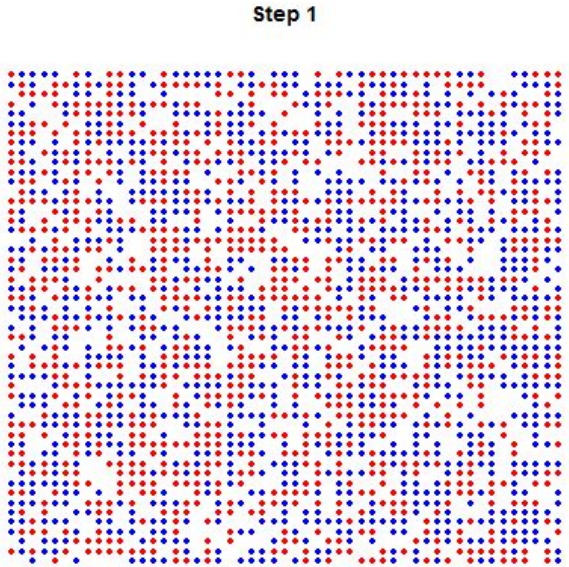
Objectives Limitation: Micromotives / Macrobehavior (2/2)

Alignment of components does not mean the whole system is aligned

- *collective action problems*: cost for a player to contribute, but others receive a benefit

Alignment at multiple levels - depending on connectivity, interactions between agents might matter more than how they act in isolation

Objectives Limitation: Micromotives != Macrobehavior (2/2)



Systemic Safety and a Leviathan

A **Leviathan** is a collective of agents which regulate bad behavior from other elements of society—prevents agents from behaving selfishly at the expense of others

Modern humans engage in the opposite:
reverse dominance hierarchy

Institutions and infrastructure for safely steering AI early.

- The infrastructure of the internet wasn't set up with safety in mind, and has long term financial and security costs.

Training Goes Awry vs Evolution

Training Objective View

- ◇ alignment with base objective is what we need
- ◇ fanatical optimizer destroys us
- ◇ dangerous AI agents as idiot savants
- ◇ singleton
- ◇ a paperclip maximizer is a dangerous agent
- ◇ any amount of misalignment results in doom
- ◇ “solving” alignment with a monolithic airtight solution
- ◇ an instrumental incentive will go to infinity
- ◇ maximizers/instrumental incentives are dangerous
- ◇ prevent humans from being suddenly wiped out

Evolutionary View

- ◇ objective is not the only thing shaping AI
- ◇ Darwinian forces erode us
- ◇ dangerous AI agents as selfish or an invasive species
- ◇ multiagent
- ◇ a fitness maximizer is a dangerous agent
- ◇ some amount of misalignment is inevitable
- ◇ domestication
- ◇ behaviors must be balanced to improve fit
- ◇ natural selection is dangerous
- ◇ prevent Darwinism from bringing us and AIs to a bad local optimum

Conclusion

An individual agent will not necessarily shore up power

How were humans domesticated? Reverse dominance hierarchies and their conscience

Need multiple levels