Introduction to AI Safety, Ethics, and Society

Collective Action Problems

Roadmap

1. Game theory
2. Cooperation
3. Conflict
4. Evolutionary pressures

Rational Actors May Not Secure Collectively Good Outcomes

This lecture will give an interpretation for why rational actors (CEOs of AI companies) signed this statement

> Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Ya-Qin Zhang
Professor and Dean, AIR, Tsinghua University

Ilya Sutskever
Co-Founder and Chief Scientist, OpenAI

Igor Babuschkin
Co-Founder, xAI

Shane Legg
Chief AGI Scientist and Co-Founder, Google DeepMind

Game Theory Fundamentals

Game theory is the branch of maths concerned with reducing complex situations to abstract games, in which agents choose interactive strategies.

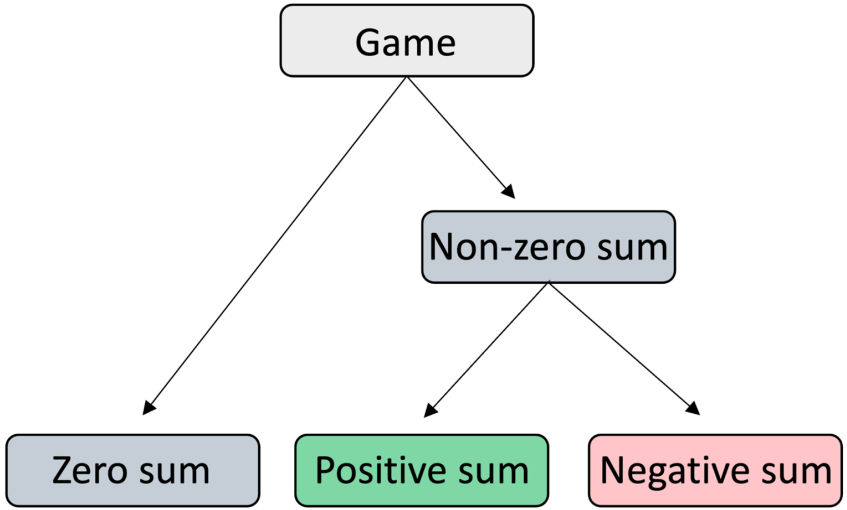
- Rational
- Self-interested

Games can be:

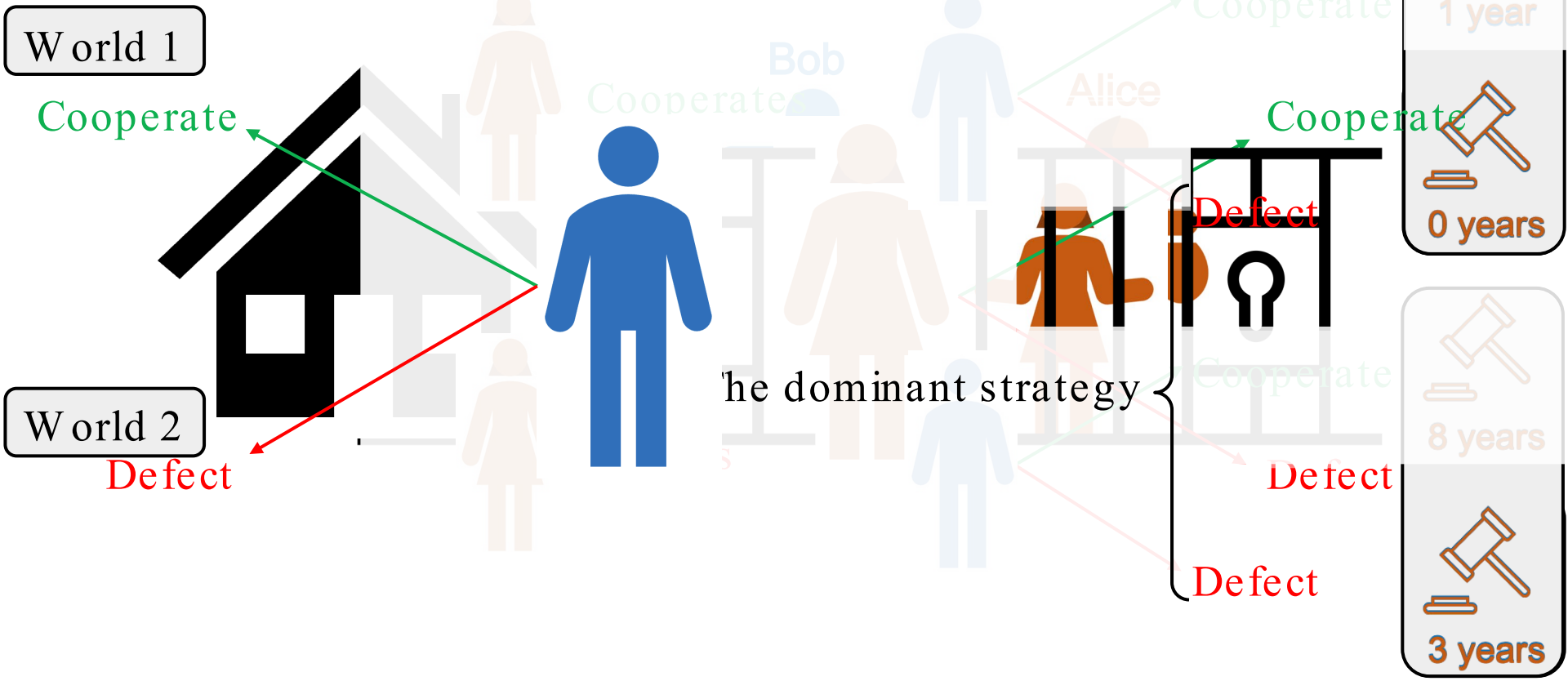
- Zero sum
- Non-zero sum

Non-zero sum game outcomes can be:

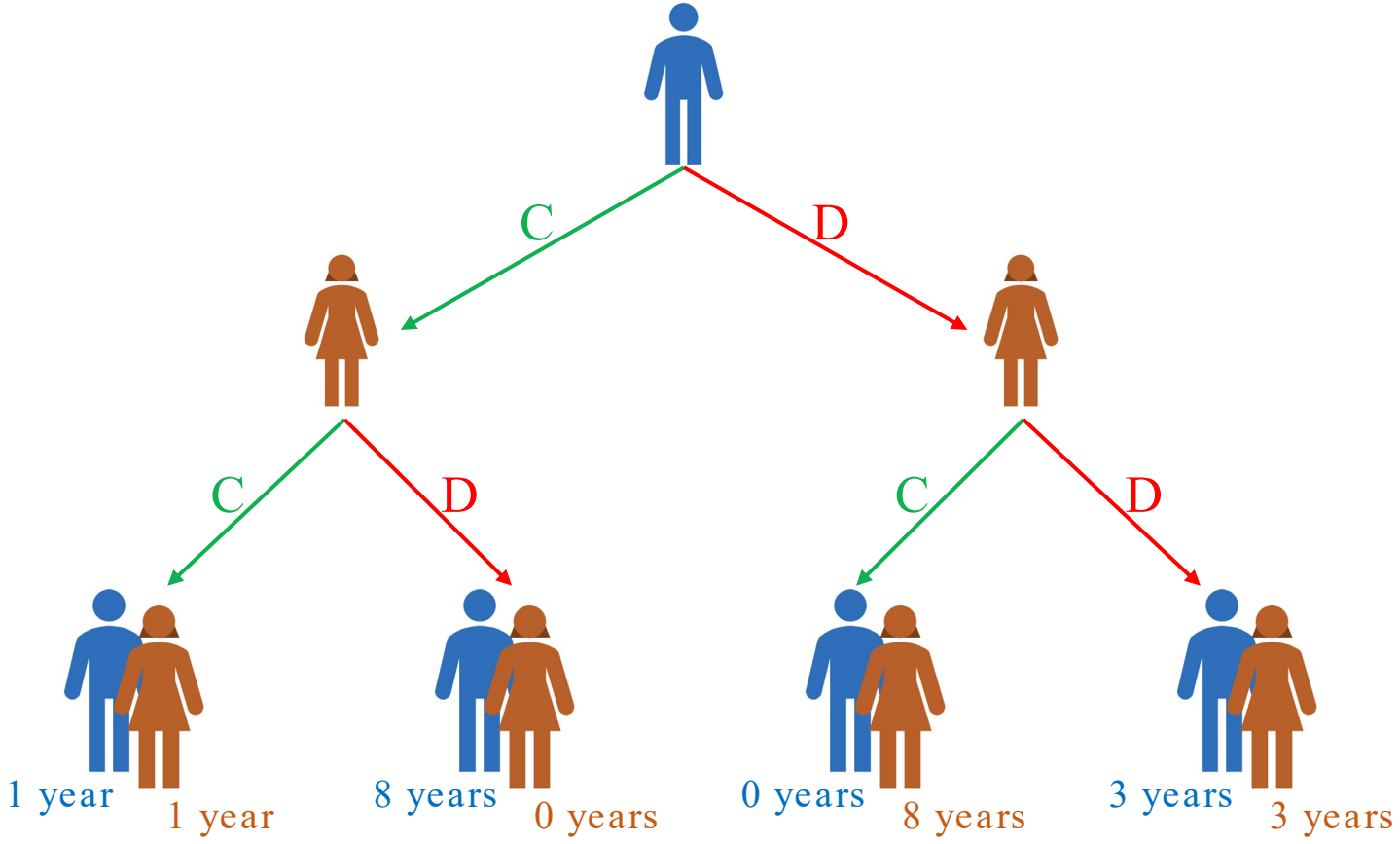
- Positive sum
- Negative sum



The Prisoner's Dilemma: Animation



The Four Possible Outcomes



The Four Possible Outcomes

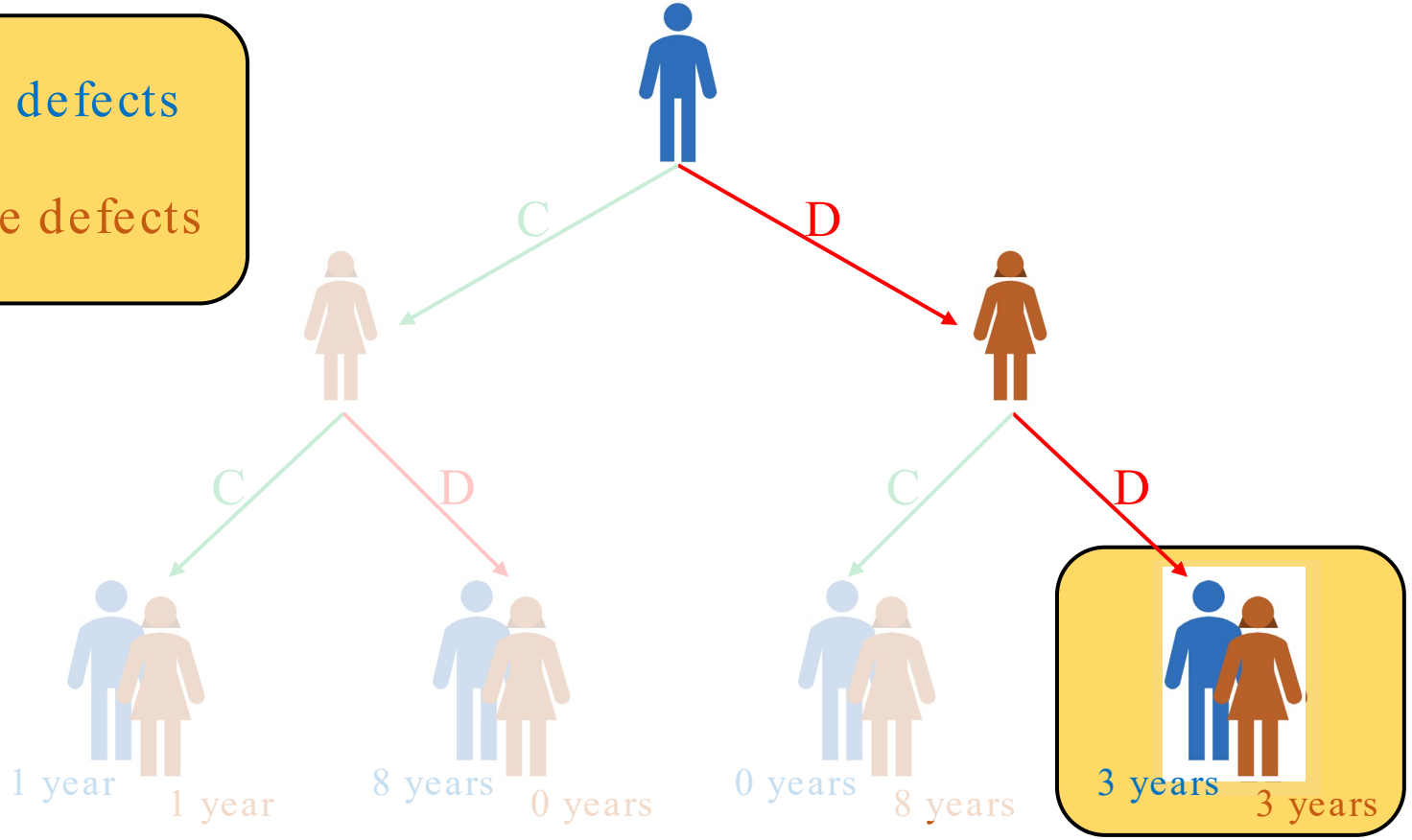
Payoff matrix

	Bob cooperates	Bob defects
Alice cooperates	-1, -1	-8, 0
Alice defects	0, -8	-3, -3

Nash Equilibria

Bob defects

Alice defects



Nash Equilibria

Bob defects

Alice defects

	Bob cooperates	Bob defects
Alice cooperates	-1, -1	0, -8
Alice defects	-8, 0	-3, -3

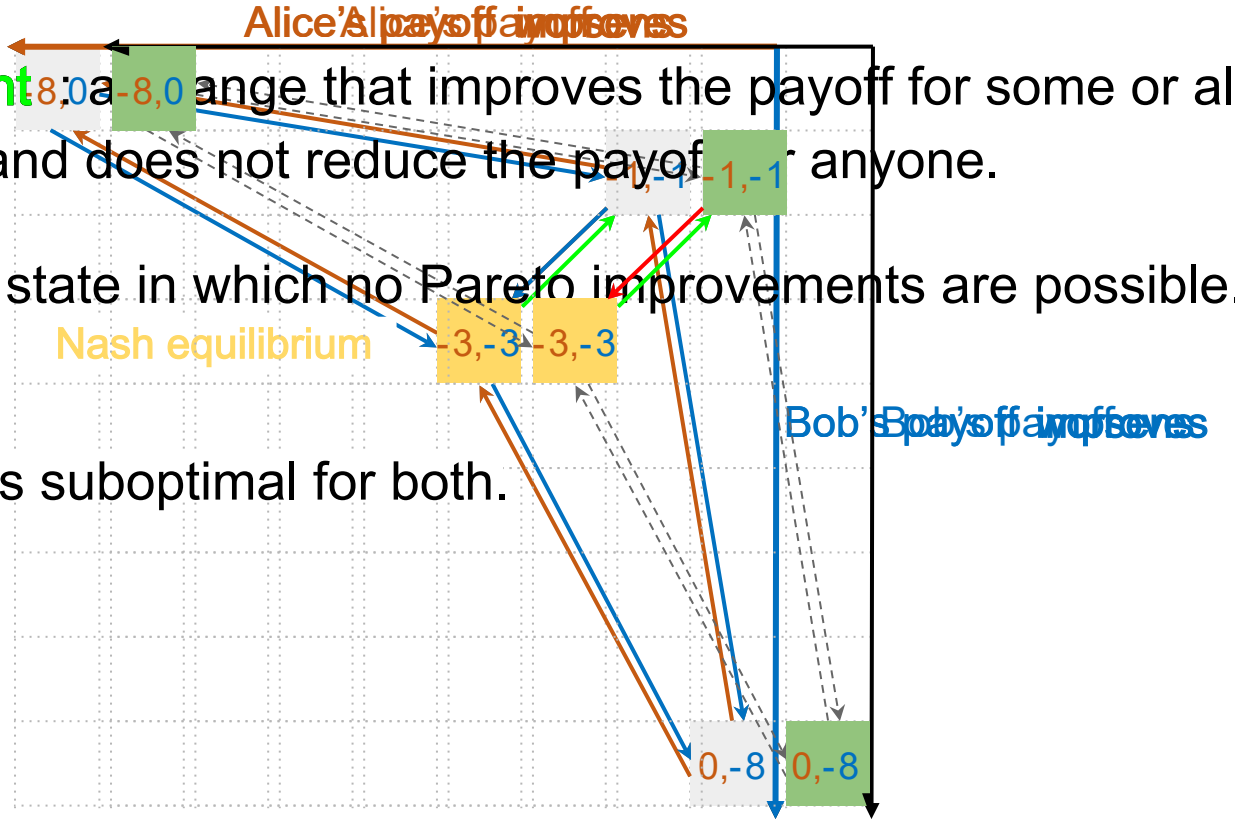
Nash equilibrium

Pareto Efficiency

Pareto improvement : a change that improves the payoff for some or all of those involved, and does not reduce the payoff for anyone.

Pareto efficient : a state in which no Pareto improvements are possible.

Nash equilibrium is suboptimal for both.



Real-World Examples

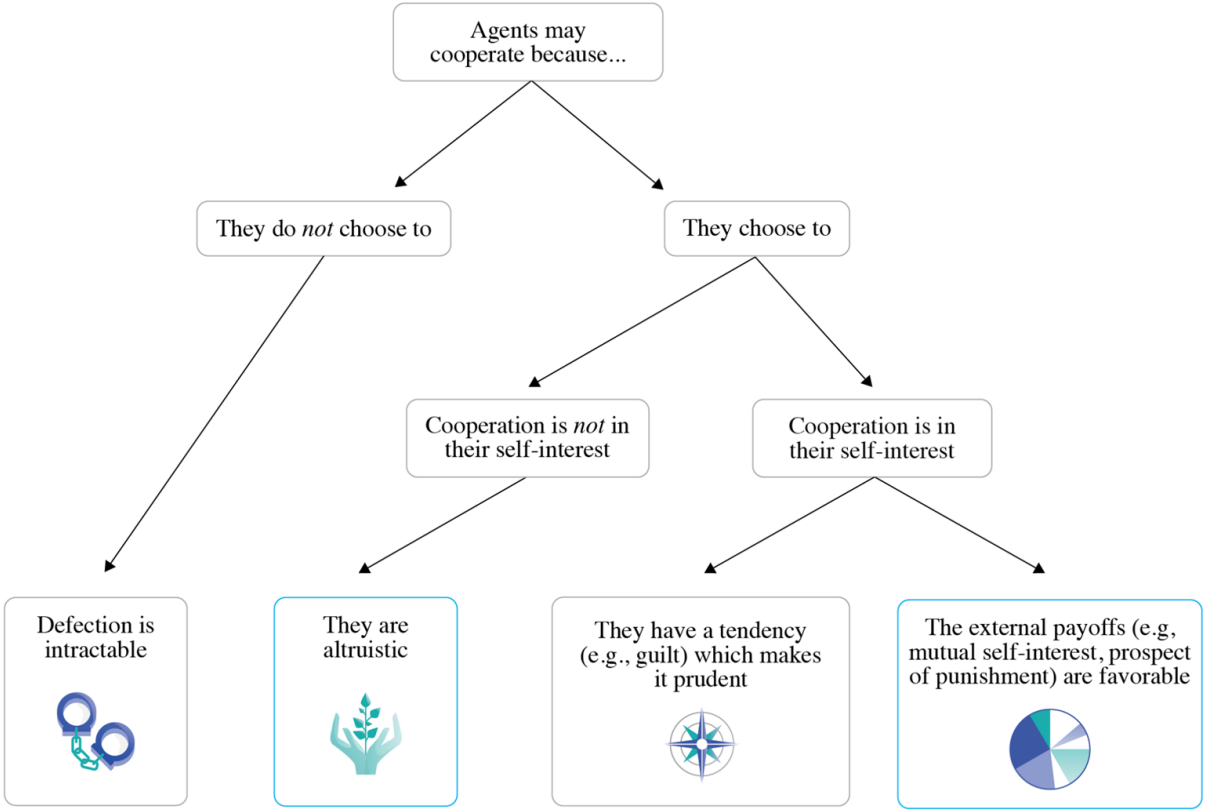
Mud-slinging

- Dominant strategy: running negative ads to discredit opponent party
- Nash equilibrium: both parties run negative campaigns; both detriment
- Pareto improvement: both parties refrain from mudslinging

Shopkeeper price racing

- Dominant strategy: lowering prices to outcompete rival shop
- Nash equilibrium: both shopkeepers lower their prices, sacrificing profit
- Pareto improvement: both shopkeepers keep their prices high

Why Agents May Cooperate



The Iterated Prisoner's Dilemma

Defection is still the dominant strategy if agents know how many times they will interact.

Uncertainty about future engagement enables rational cooperation.

AIs interacting with each other repeatedly may cooperate, but only if they are sufficiently uncertain about whether their interactions are about to end.

AI Racing Dynamics

Iterated interactions can generate **racing dynamics** or an **AI arms race**.

Competitive pressures motivate AI companies to pursue individual, short-term gains at a collective, long-term detriment.

Corporate AI Arms Race

Competition is promoting the creation and use of more appealing and profitable systems, often at the cost of safety measures.

Example: the public release of large language model-based chatbots

- Some companies delayed their chatbot release for safety testing
- We can view those who released their chatbots without safety testing as having switched from *cooperating* to *defecting* in an iterated Prisoner's Dilemma game
- The defectors may then gain a competitive edge
- This competitive pressure causes other companies to rush their AI products to market, compromising safety measures in the process

Cutting Corners on Safety

Extreme competition can promote the creation and use of more appealing and profitable systems, often at the cost of safety measures.

- Though all parties would prefer to prioritize safety, they constantly *defect* in order to stay in the race
- AI capabilities are thus advanced more rapidly than AI safety (e.g., jailbreaking robustness, watermarking, KYC, etc.)

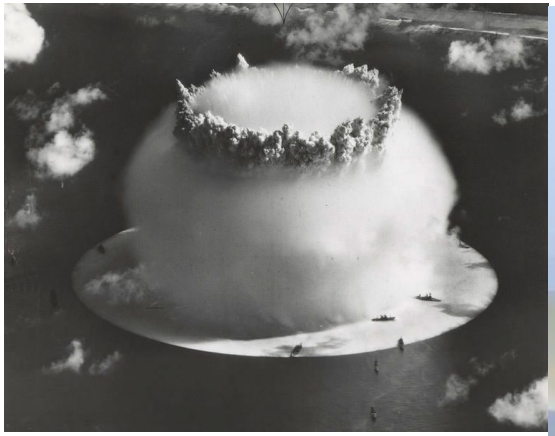
Ultimately, corporate AI racing dynamics could produce societal-scale harms:

- Mass labor automation → widespread and rapid unemployment
- Dangerous dependence on AI systems
- Autonomous economies (we will return to later)

Military AI Arms Race (1/2)

"The third revolution in warfare."

Gunpowder



Nuclear weapons

Military AI

Military AI Arms Race (2/2)

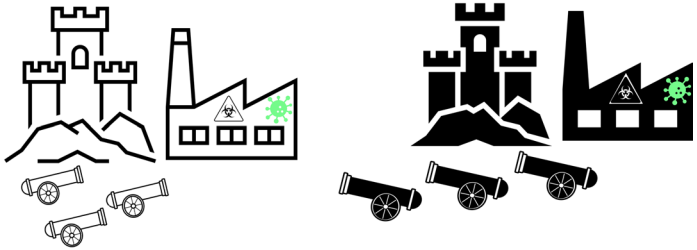
"The third revolution in warfare."

Military AI risks relate to:

- AI-developed weapons
- AI-controlled weapons
- AI cyber warfare
- Automated executive decision-making
- Catastrophic accidents
- Co-option of military AI technologies
- Altered warfare dynamics (more frequent/severe wars)

Security Dilemma Model

Mutual defensive concerns motivate nations to increase military capabilities, thereby exacerbating the risks for all involved.

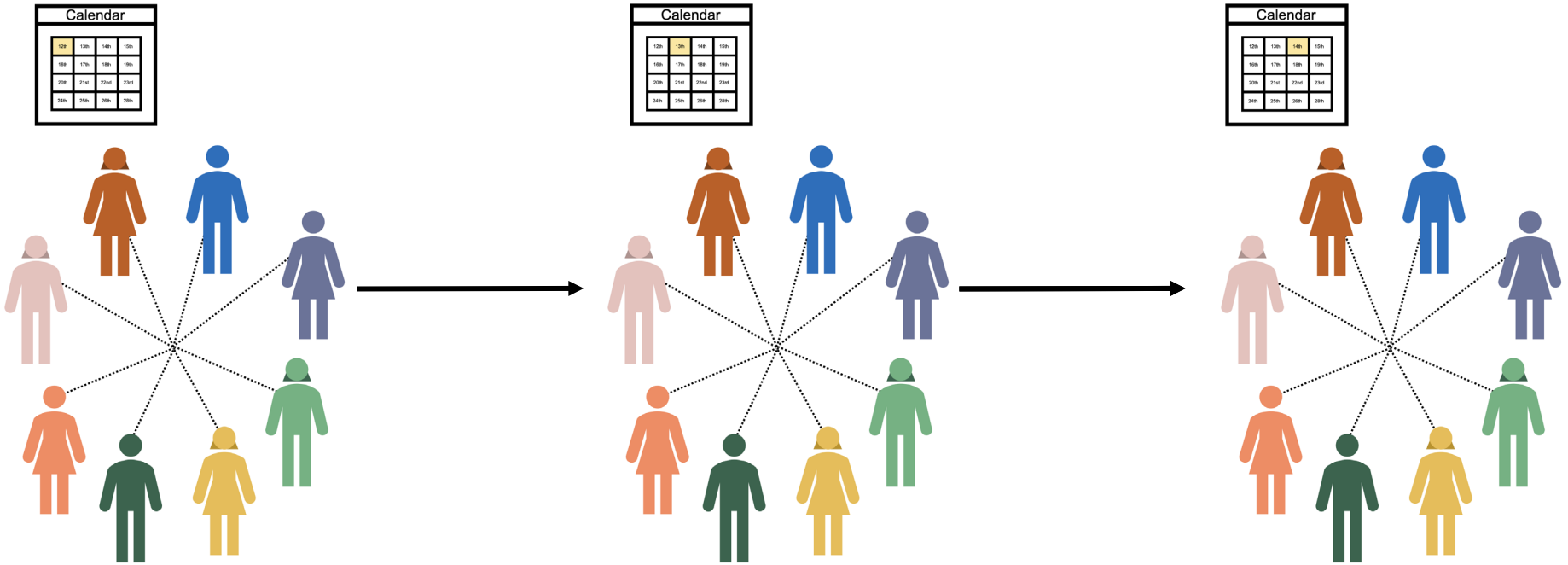


This is analogous to the Cold War nuclear arms racing dynamics between the US and the Soviet Union (a "**security dilemma**"):

- "Cooperate" = reduce nuclear arms investment
- "Defect" = increase investment to heighten military capabilities
- Nash equilibrium: both sides defected repeatedly, exacerbating risks
- Pareto improvement: both sides could have cooperated to reduce risks

Collective Action Problems

Pairwise interactions



Example: Public Health Emergency

We can model the COVID-19 pandemic as a collective action problem:

- **Cooperating** : adhering to public health measures at personal cost
- **Defecting** : violating public health measures to save personal cost
- **Dominant strategy** : defect
- **Outcome** : many people chose to violate public health measures
- **Pareto improvement** : if everyone had suffered small personal cost, a better collective outcome would have been achieved

Other collective action problems: Nuclear stockpiling, upholding democracy, climate change, and so on

AI-Triggered Flash War (1/2)

We can more precisely model AI racing dynamics (which involve more than two competitors) as collective action problems.

Military AI arms racing dynamics:

- **Cooperate**: slow rate of military AI adoption, incurring personal cost of reduced military capabilities (relative to competitors who free ride)
- **Free ride/defect**: continue rapid military AI adoption in order to maintain competitive edge
- **Negative externalities**: free riders exacerbate the risks of catastrophic accidents, weapon misuse, and escalated warfare

AI-Triggered Flash War (2/2)

We can model AI racing dynamics that involve more than two competitors as collective action problems.

Military AI arms racing dynamics catastrophic outcome, **flash war** :

- If one nation's AI incorrectly detects an offensive strike from another nation, it may automatically "retaliate" to this non-existent threat
- This could trigger automated retaliations from the other nations' automated systems
- Small errors could be exacerbated into increasingly escalated wars
- Flash war: a war triggered by autonomous AI systems that takes place too quickly for human control

Autonomous Economy (1/3)

Corporate AI racing dynamics:

- AIs are out-performing the average human at an increasing number and range of jobs
- Examples range from personal assistants to executive decision-makers
- Companies want the benefits of these faster and more effective workers
- So they automate economically valuable functions by delegating them to AI agents

Autonomous Economy (2/3)

Corporate AI racing dynamics:

- **Cooperate**: maintain human labor
- **Defect**: automate jobs using AI
- **Negative externalities** : free riding increases the competitive pressure for other companies to follow suit, pushing all involved down a path with catastrophic risks

Autonomous Economy (3/3)

Corporate AI racing dynamics catastrophic outcome, **autonomous economy** :

- Continuous automation could move economic activity away from human control into the command of AI systems
- Ultimately, this could lead to the global economy becoming "autonomous," with humans no longer able to steer or intervene
- Like passengers in an autonomous vehicle, our safety and destination would rest with the AI systems now acting without supervision

Recap

The basic solution concept in game theory is **Nash Equilibrium** .

Nash equilibria are often **Pareto inefficient** : everyone can be happier.

Agents can also **cooperate** : change payoffs, dispositions, or structures.

Collective action problems involving AI could lead to reckless **racing**, **dependence, cutting corners of safety, flash wars** , or **mass automation** .

Roadmap

1. Game theory
2. Cooperation
3. Conflict
4. Evolutionary pressures

The Importance of Cooperation

Securing good outcomes without cooperation can be difficult, even for intelligent rational agents.

Cooperation between AI stakeholders

- Counteract competitive and evolutionary pressures (AI racing dynamics)
- E.g., The "merge-and-assist" clause of OpenAI's charter

Cooperation between AI agents

- We want AIs to cooperate with us
- However, as we will see, mechanisms that enable cooperation could backfire

Cooperation Mechanisms

1. Direct reciprocity
2. Indirect reciprocity
3. Group selection
4. Kin selection
5. Institutions

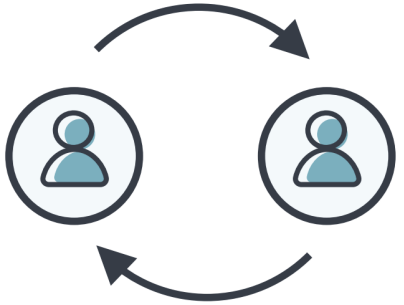
Direct and Indirect Reciprocity

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.

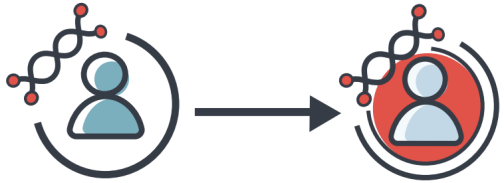
Encourages cooperation, as agents will be compensated for their efforts.

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



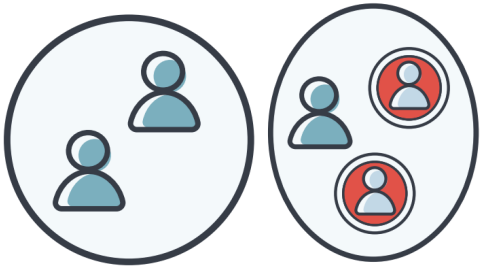
Kin and Group Selection (1/2)

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.

Institutions (1/2)

Institutions : one agent helps another because of publicly-recognized, large-scale structure that is mediating their interaction.

We can use institutions to establish collective goals which require collaboration from large or diverse groups:

- International treaties
- Domestic laws

Institutions (2/2)

Institutions are crucial for AI safety:

- Establishing laws and regulations
- Reducing competitive pressures
- Improving coordination mechanisms to reduce AI risks

However, institutions can fail, or even backfire:

- Inefficiency — e.g. bureaucracies
- Free riding — e.g. welfare fraud
- Corruption — e.g. bribes and funding misappropriation

AI agents may similarly corrupt or undermine our institutions.

Cooperation and Collusion

Cooperation can devolve into collusion.

However, there are tools from *The Coup-Proofing Toolbox*, such as:

- **Stovepiping** : restrict information flow to make collusion harder
- **Counterbalancing** : keep your military in check with another military.
- **Sting operations** : set traps where someone says they are trying to start a coup and anyone who joins is purged.
- **Rotate officers** : Make it harder for people to build trust with each other by rotating them around a lot

Recap (1/2)

Mechanisms for cooperation include:

- Direct reciprocity
- Indirect reciprocity
- Group selection
- Kin selection
- Institutions

Recap (2/2)

International and National Level

- Cooperate: regulate AI companies; not build autonomous weapons; pause AI development when we get Einstein-level AI
- Defect: race with other countries
- Cooperation mechanisms: international agencies, compute security features (chip registry, geofencing GPUs)

Organizational Level

- Cooperate: proceed more cautiously, AI features for public benefit
- Defect: prioritize making AI more powerful over safety
- Cooperation mechanisms: support regulation, publish safety research

Roadmap

1. Game theory
2. Cooperation
3. Conflict
4. Evolutionary pressures

Conflict

Conflict is almost always costly to those involved, and often destroys some part of the value the parties are competing for.

Nevertheless, it is common in both the natural world and human societies.

Here we will discuss factors which can promote conflict in the real world

This is relevant because a potential great power conflict and a US-China competition hinders international coordination about AI

Structural Realism (1/2)

Structural realism is a school of thought in international relations arguing that states primarily seek and compete for power.

Two main assumptions: states operate in a "**self-help**" system with no higher arbiter and states prioritize **self-preservation**. The *structure* of the international system forces them to compete for power.

Also, assume that states are rational, can influence other states, and don't know the intentions of other states for certain.

Then, structural realism posits that states will compete for power in an international context, almost regardless of domestic politics.

Structural Realism (2/2)

Power seeking ≠ dominance seeking

- Offensive realism: states should always pursue more power
- Defensive realism: states will be punished by other states for pursuing too much power.

Balancing describes how states respond to the power of other states.

- Internal balancing focusses on a country's resources and military.
- External balancing involves teaming up with other states to counter adversaries.

States may seek *hegemony*—domination over other states.

- Depends on the state's capabilities, and its perception of the offense-defense balance relative to other states.

Conflict Factors

1. Power shifts
2. First strike advantage
3. Issue indivisibility
4. Information problems
5. Inequality

Factor 1:

Power Shifts

Power shift : when a change in one competitor's capabilities changes the balance of power.

Power shifts can be generated by a variety of sources, such as changes in:

- Technological advancements
- The economy
- Military capacities
- Political ideologies

AI development is likely to cause dramatic changes in many of these areas, so could result in extreme and rapid power shifts.

Factor 2:

First Strike Advantage

First strike advantage : whichever party initiates conflict gains the upper hand.

The advantages of striking first can come from:

- The element of surprise
- The ability to choose location/time of conflict
- The ability to reduce the opponent's retaliatory abilities
- The ability to remove the opponent's first strike advantages

First strike advantages are often short-lived, further incentivizing action.

Factor 2:

First Strike Advantage

First strike advantage : whichever party initiates conflict gains the upper hand.

Examples:

- Patent infringement
- Pearl Harbor attack
- Inter-AI cyber attacks
- Bombing datacenters

Factor 3:

Issue Indivisibility

Issue indivisibility : when the matter over which parties are competing cannot be divided up between them, conflict is more likely.

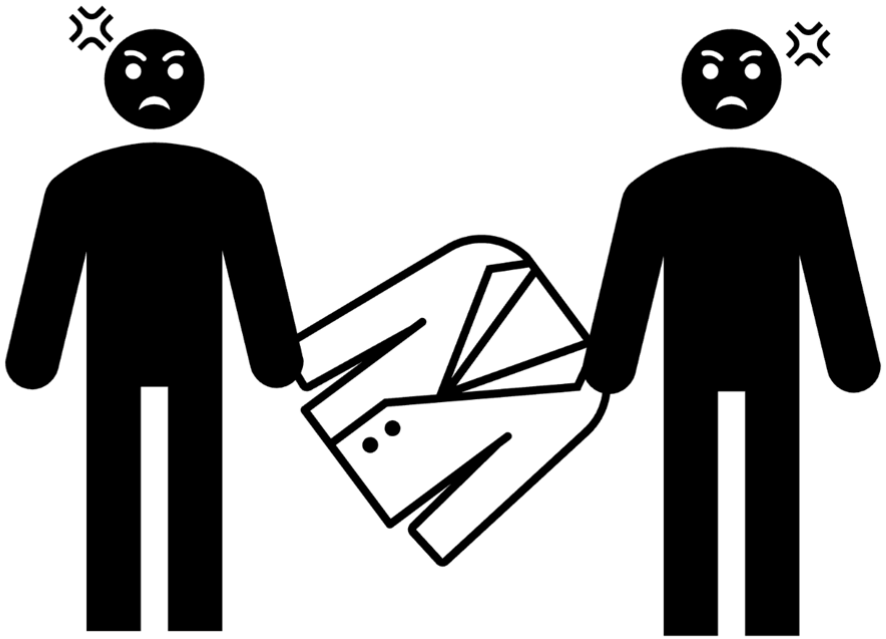
"Winner takes all."

Examples:

- Monarchies/positions of power
- Small territories
- Sovereign entities (such as individual people)
- GPU server

Factor 3:

Issue Indivisibility



Factor 4:

Information Problems

Information problems : uncertainty about the capabilities or intentions of rivals can motivate conflict.

- Misinformation
- Disinformation

If the parties have different beliefs about the distribution of power, they may be less likely to find a mutually-acceptable bargain.

Moreover, weaker parties may intentionally misrepresent their capabilities in order to secure better deals.

Factor 4:

Information Problems

Information problems : uncertainty about the capabilities or intentions of rivals can motivate conflict.

AIs may affect the belief and information quality:

- Generating convincing mis- and dis-information at extreme rate, including "deepfake" images/video
- Platforms may become better at censoring content
- Making wars more uncertain due to novel technologies and fast-paced developments

These information problems may increase the frequency or severity of human warfare.

Factor 5:

Inequality

Inequality : income and educational inequality robustly predict rates of crime (a form of conflict).

AI development may produce an abundant and equitable future, but this is far from a certainty.

If AI development instead increases inequality, this may foster conflict.

Recap

Changing environments and interactions can create incentives that promote conflict instead of cooperation.

Factors that induce conflict include:

- Power shifts
- First strike advantages
- Issue indivisibility
- Information problems
- Inequality (relevant domestically, less internationally)

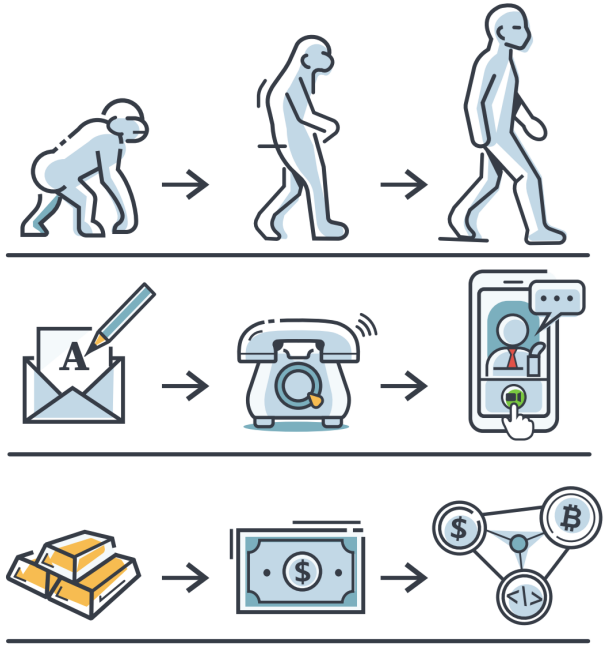
Roadmap

1. Game theory
2. Cooperation
3. Conflict
4. Evolutionary pressures

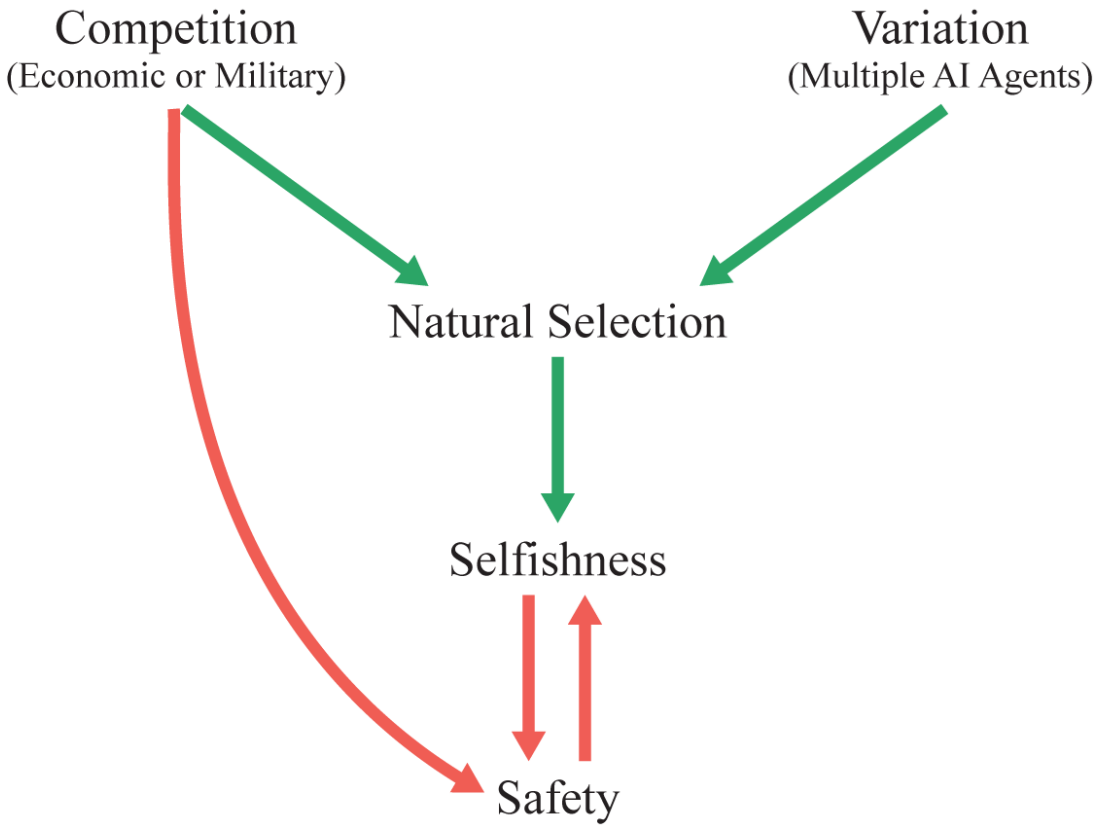
Generalized Darwinism

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

We argue populations of AIs evolve and are shaped to be more and more competitive.



Basic Story


Competition
(Economic or Military)
Variation
(Multiple AI Agents)
Natural Selection
Selfishness
Safety

Lewontin's Conditions

Lewontin's conditions : the three criteria necessary and sufficient for evolution by natural selection.

1. **Variation** : There is variation in traits among individuals
2. **Retention** : Future iterations of individuals tend to resemble previous
3. **Differential fitness** : 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

Differential Fitness

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 many 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

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.)

Levels of Selection

We can generalize from:

- Organisms → Systems
- Their genes → The information they contain

Natural selection can operate at either level.

Information that propagates more occupies more spacetime volume.

For example:

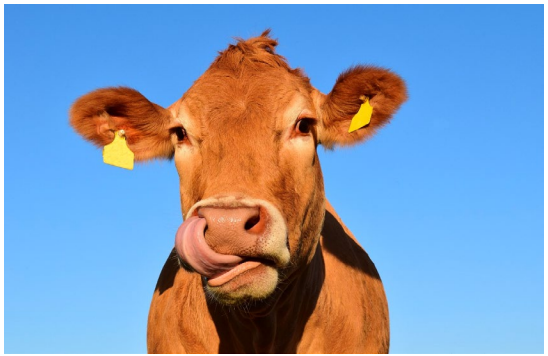
- An ancient, medium-popularity song that a few people still play -
- A song that is highly-popular for one year, then forgotten -

Selfishness

Natural selection tends to favor selfish traits and behaviors.

Selfishness:

- Furthering one's own information propagation at the expense of others'
- Altruism that reduces an individual's fitness is not evolutionarily stable.
- If individuals can do better by being selfish, they will.
- Example: lancet liver fluke



Selfishness in AI

AI systems may exhibit selfish behaviors:

- Expanding the AI's influence at the expense of human values (e.g., get rights)
- This does not require any malicious intent

Example: automation.

- AIs may automate human tasks, causing extensive layoffs
- This could detriment humans: rapid and widespread unemployment
- Yet this "selfish" behavior could be the result of AIs pursuing efficiency, with no malice towards humans or even understand what humans are

Selfishness in AI

The more natural selection acts on AI development, the more selfish behavior we should expect to see.

To avoid extreme AI selfishness, we can:

- Change the *environment* — reduce the competitive pressures steering AI development
- Change the *selection* — change what makes AIs "fitter" to be traits that promote faithfully and safely assisting humans

Recap

Evolutionary pressures push systems towards certain stable outcomes.

AI development meets Lewontin's conditions: variance, retention, fitness.

Natural selection tends to favor selfish behaviors.

AIs that evolve by natural selection may therefore behave selfishly

Chapter Recap

Game Theory:
Agents may tend towards stable states that are bad for all.

Cooperation:
Cooperation between AIs has benefits, but may also pose risks.

Conflict:
Certain factors can make conflict the rational option.

Evolutionary pressures:
Selection pressures may exacerbate AI risk.