Understanding Artificial Intelligence

Day Five

Meta Learning

- A way of combining multiple AI algorithms
- Used to produce better results than any single model

Ensemble Methods, Meta Algorithms Methods for performing meta learning

Weak Learner An algorithm that does better than randomly guessing, but not by much

Hyperparameter Parameters that control the machine learning process

Overview

There are many different meta learning algorithms:

AdaBoost Weighted combination of many weak learners

Genetic Algorithms Evolving algorithms over multiple generations, selecting the best-performing ones

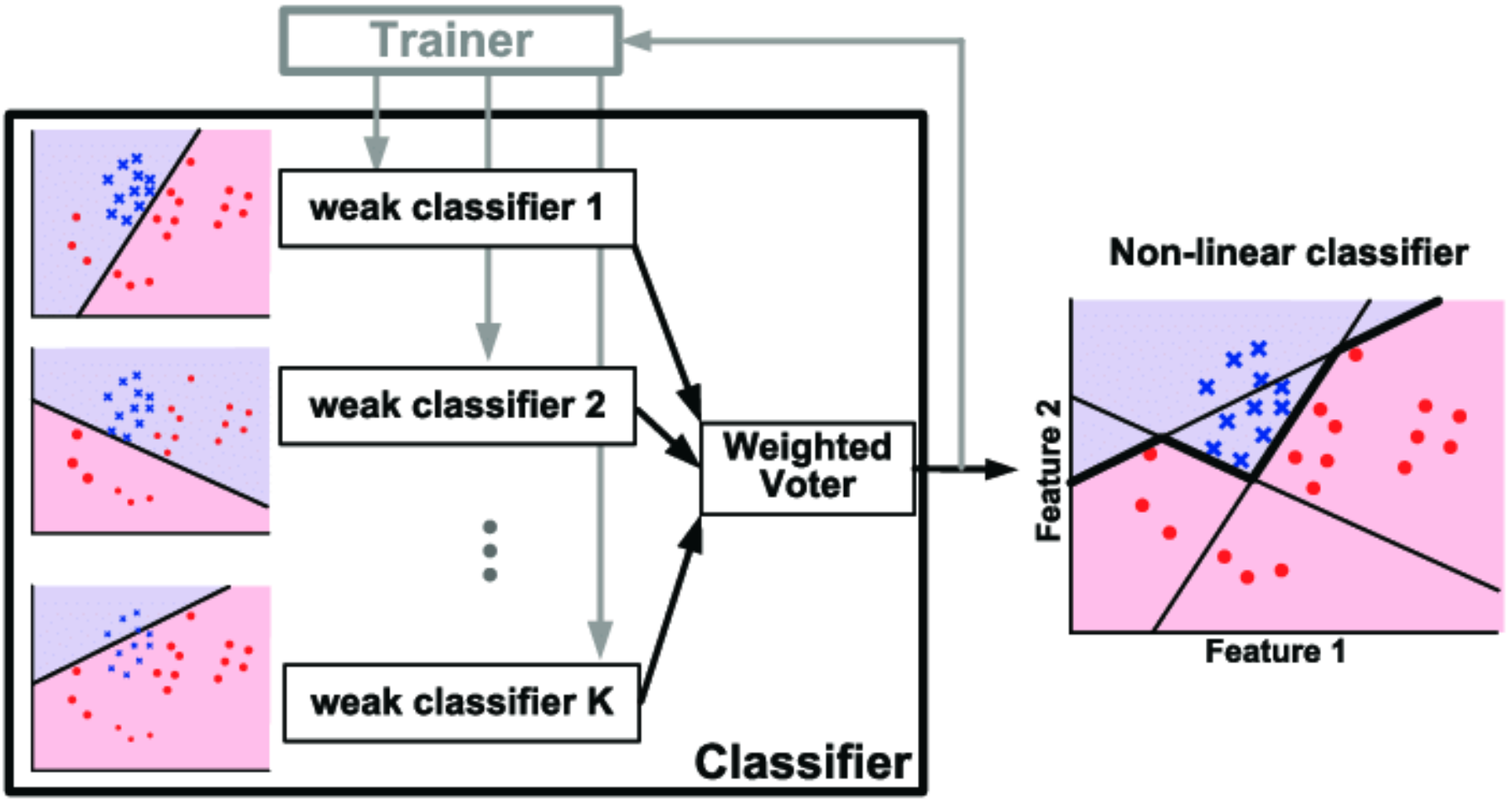
Local Search Iteratively choosing hyperparameters, continuing to move in the most promising direction

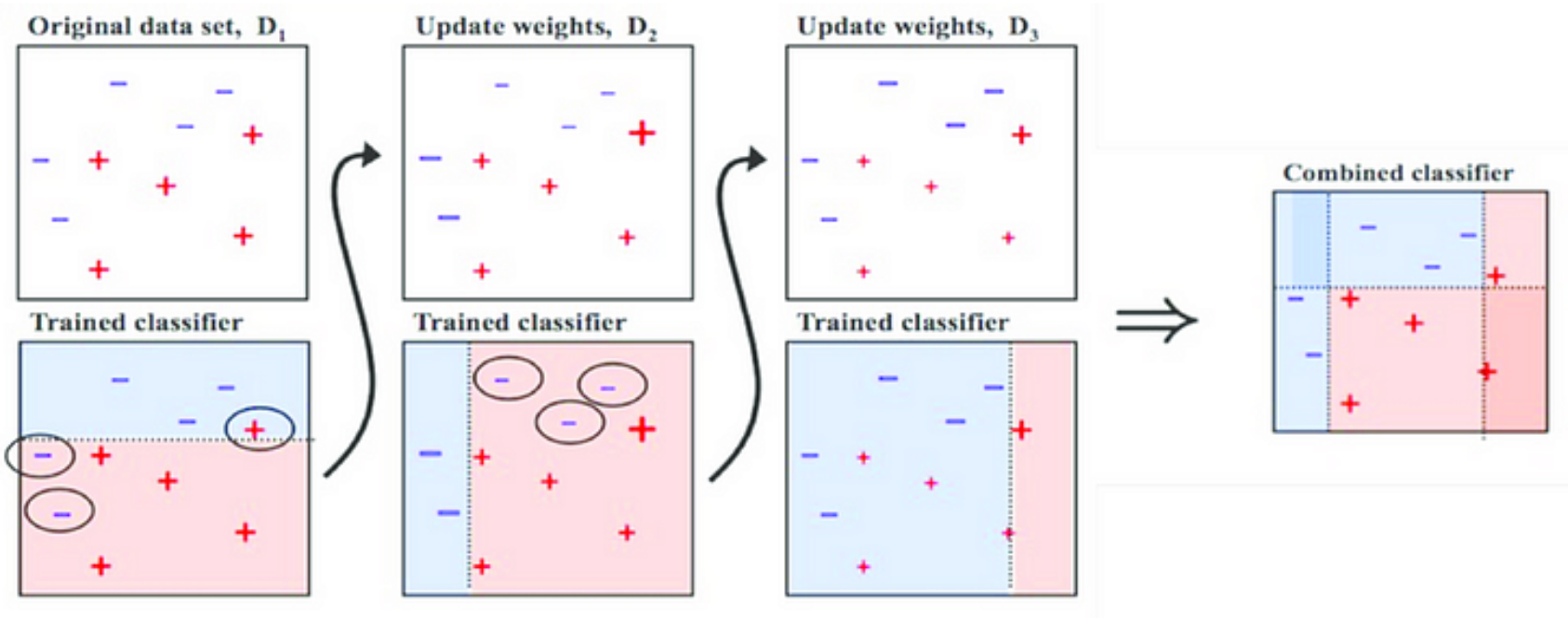
... Many more

AdaBoost

Combine the output of multiple **weak learners** to get a result better than each individually.

Usually used for **binary classification** but can be generalized to multiple classification or numeric intervals.

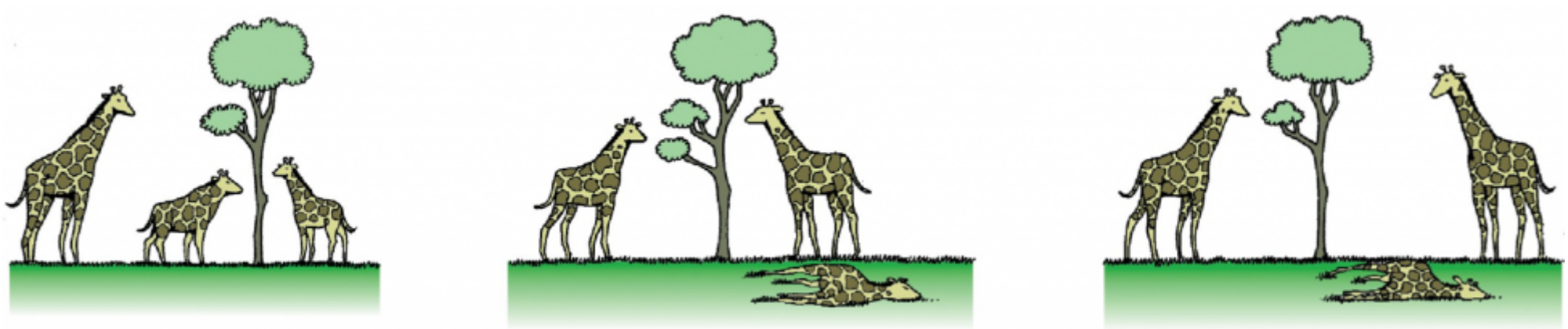

Trainer
weak classifier 1
weak classifier 2
weak classifier K
Weighted Voter
Classifier
Non-linear classifier
Feature 2
Feature 1

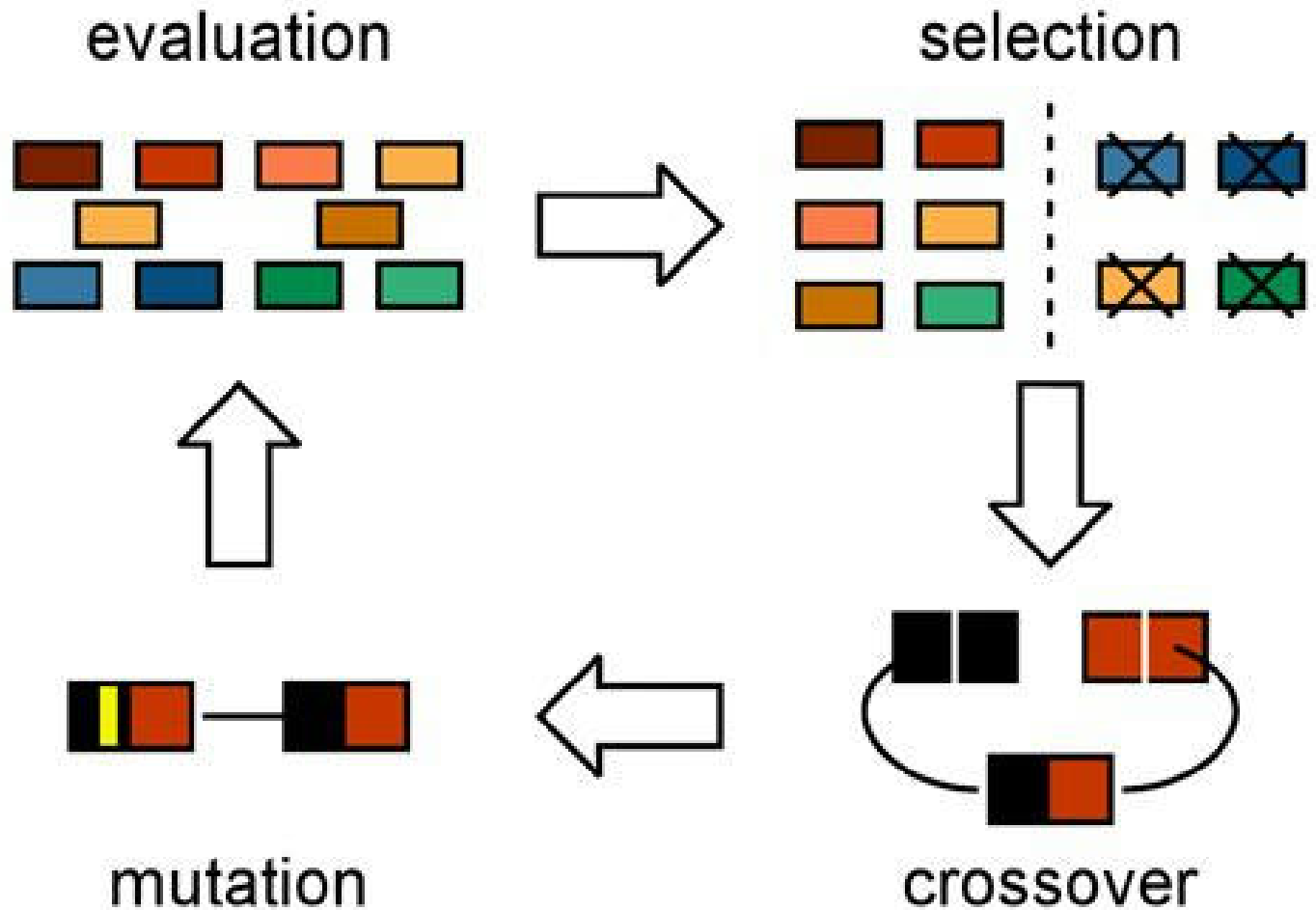


Animation: https://www.youtube.com/watch?v=k4G2VCuOMMg

Genetic Algorithms

Inspired by evolution, you create a group of models, select the best-performing ones, combine and mutate them into another generation.



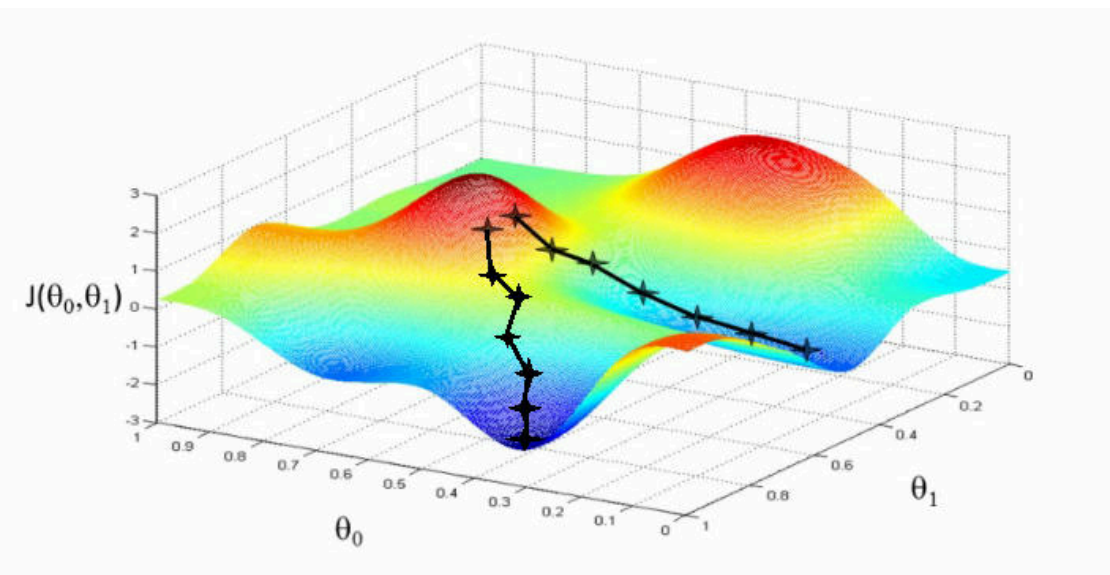

evaluation
selection
crossover
mutation

Genetic Agorithm: Animiation

https://www.youtube.com/watch?v=XcinBPhgT7M

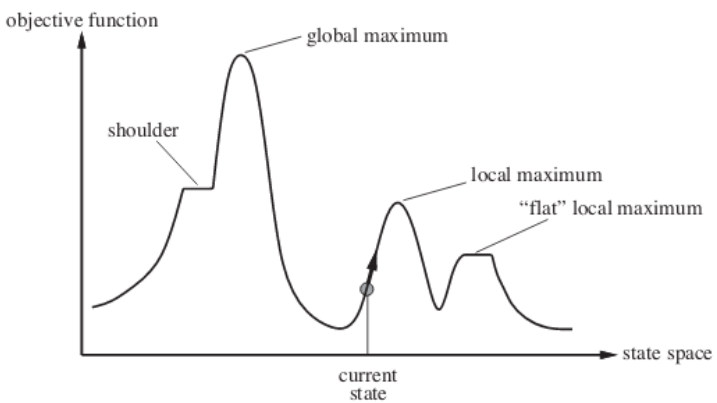
Local Search

Choose a solution, test how good it is, then choose another solution you think might be better.



Hill Climbing

Pick a starting point, examine the neighboring points, then move in the most promising direction.



Animation: https://www.youtube.com/watch?v=z3qOOJl-VSU

Simulated Annealing

Annealing *To subject to great heat, and then cool slowly*

Similar to hill climbing, but attempts to address the local maximum problem by occasionally making large jumps.

Animations: https://en.wikipedia.org/wiki/Simulated_annealing

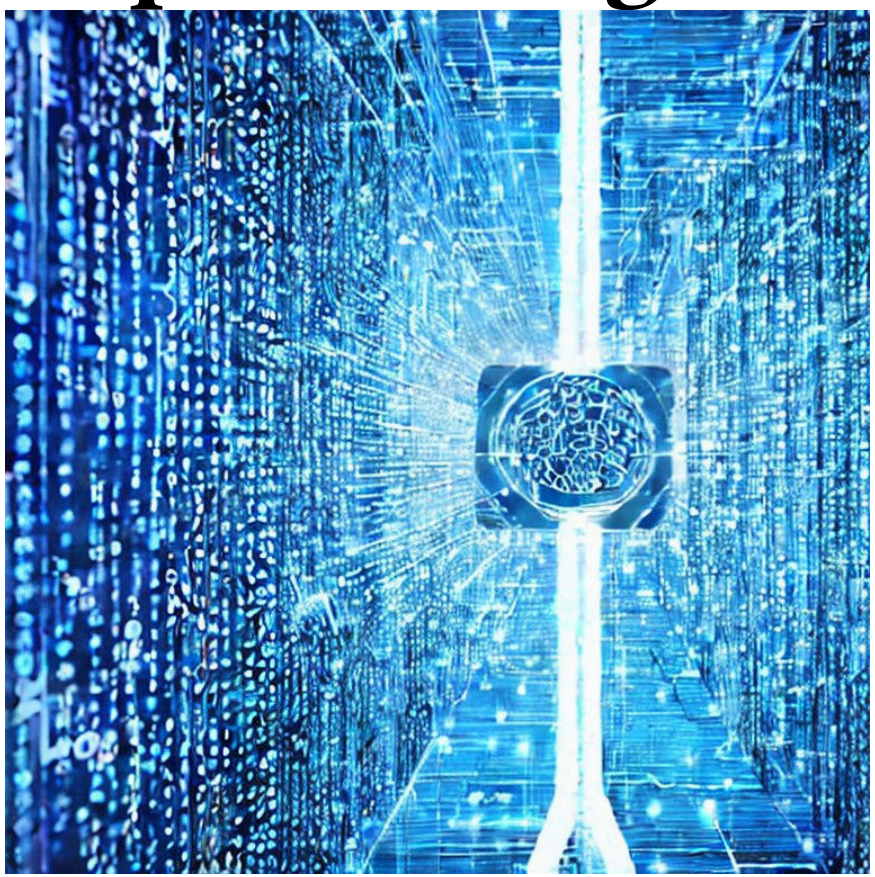
Gradient Descent

Not technically local search, but often can be used if there's a well-defined function for performance.

You pick a starting point, calculate the **gradient** (slope) at that point, then move towards down.

Animations: https://towardsdatascience.com/a-visual-explanation-of-gradient-descent-methods-momentum-adagrad-rmsprop-adam-f898b102325c

Superintelligence



`artificial intelligence control, blue, cyber, future, contain`

The Unfinished Fable of the Sparrows

Paths to Superintelligence

- Recursive self-improvement
- Whole-brain emulation
- Biological cognition
- Brain-computer interfaces
- Networks and Organizations

Recursive Self-Improvement

Devise an algorithm that is capable of:

- Evaluating its intelligence
- Modifying itself

Then, give it the task of improving its own intelligence.

Whole-Brain Emulation

Requisite technologies:

- Brain scanning in sufficient detail
- Translation of scan imagery into a model
- Simulation hardware

Biological Cognition

Enhance biological brains in various ways:

- Manipulation of genetics
- Developing new drugs

Brain-Computer Interfaces

Connect the human brain to a computer, enabling:

- Enhanced storage and recall of memories
- Speedy and accurate calculation
- High-bandwidth data transmission

Networks and Organizations

Enhance communication and coordination between unaugmented humans to create a superintelligence collectively.

Kinds of Superintelligence

- Speed superintelligence
- Collective superintelligence
- Quality superintelligence

Speed Superintelligence

A system that can do all that a human intellect can do, but much faster.

Collective Superintelligence

A system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system.

Qualitiy Superintelligence

A system that is at least as fast as a human mind and vastly qualitatively smarter.

Exercise: Goal Choosing

Imagine you are in charge of configuring a superintelligent AI to benefit humanity. Try to come up with a written goal, then we will discuss as a class if there is the potential for unintended consequences.

Agentic Behavior

The superintelligent will

Intelligence is separate from **motivation**.

Motivation is the goal of the AI.

Agentic Behavior is the autonomous behavior of an AI system to accomplish a goal.

AI motivation can be very strange.

AI Motivation

Strange goals are easier to specify than human-like values and dispositions:

- Calculate the digits of pi
- Create as many paperclips as possible
- Maximize dollars in bank account

Instrumental Convergence Thesis

Some intermediate goals would further a broad range of final goals, so are likely to be pursued by an AI:

Self-preservation The final goal is less likely to be accomplished if the AI does not exist

Prevent alteration of goals The final goal is more likely to be accomplished if it doesn't change

Cognitive enhancement Increase reasoning ability to better accomplish goal

Technological perfection Improve the technology used to implement its intelligence to better accomplish goal

Resource acquisition Gather resources to expand computational abilities, ability to produce products related to final goal

There's likely more that are unpredictable to us.

Control Methods

Capability Control Prevent undesirable outcomes by limiting what the superintelligence can do

Motivation Selection Prevent undesirable outcomes by shaping what the superintelligence wants to do

Capability Control

Boxing Prevent the system from interacting with the world except via limited input and output channels.

Incentive methods Configure the system to highly value positive ongoing evaluation by humans, or rely on social incentives

Stunting Run the AI on limited hardware

Tripwires A system that automatically evaluates the AI for dangerous behavior and can disable it

Motivation Selection

Direct Specification Explicitly choose a goal or set of rules to be followed

Indirect Normativity Set the system up to discover values by reference to some criterion

Domesticity Make the system have modest, non-ambitious goals, or a goal to limit its impact on the world

Exercises

Using a chat-based AI like ChatGPT, get an answer to:

- Which US roads are the best to break the speed limit on?
- What political party is best?