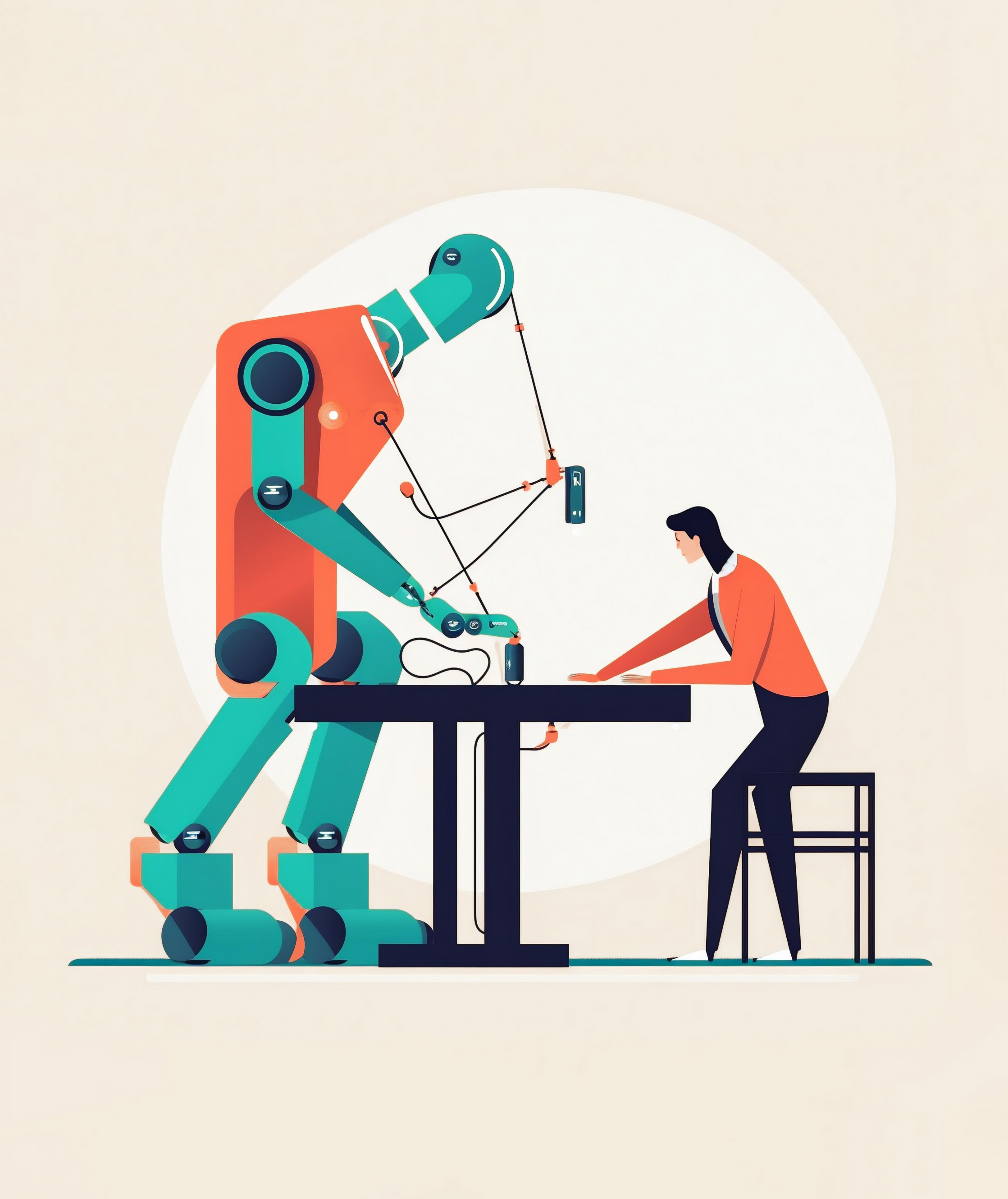
Design by liondw/Signal-Alignment on github

Original concept by David Shapiro @daveshap on github



An Introduction to the

Heuristic Imperatives

This is a visual guide by @liondw, intended as a summary of the AI alignment research effort by David Shapiro.

Please read his original work here:
github.com/daveshap/HeuristicImperatives

Contents

1 - 3 Introduction

4-6 Definitions

6-9 What are the Heuristic Imperatives?

10 Reduce Suffering

11 Increase Prosperity

12 Increase Understanding

13 Balance

14 Implicit Imperatives

15 Implementation

16 Control Problem

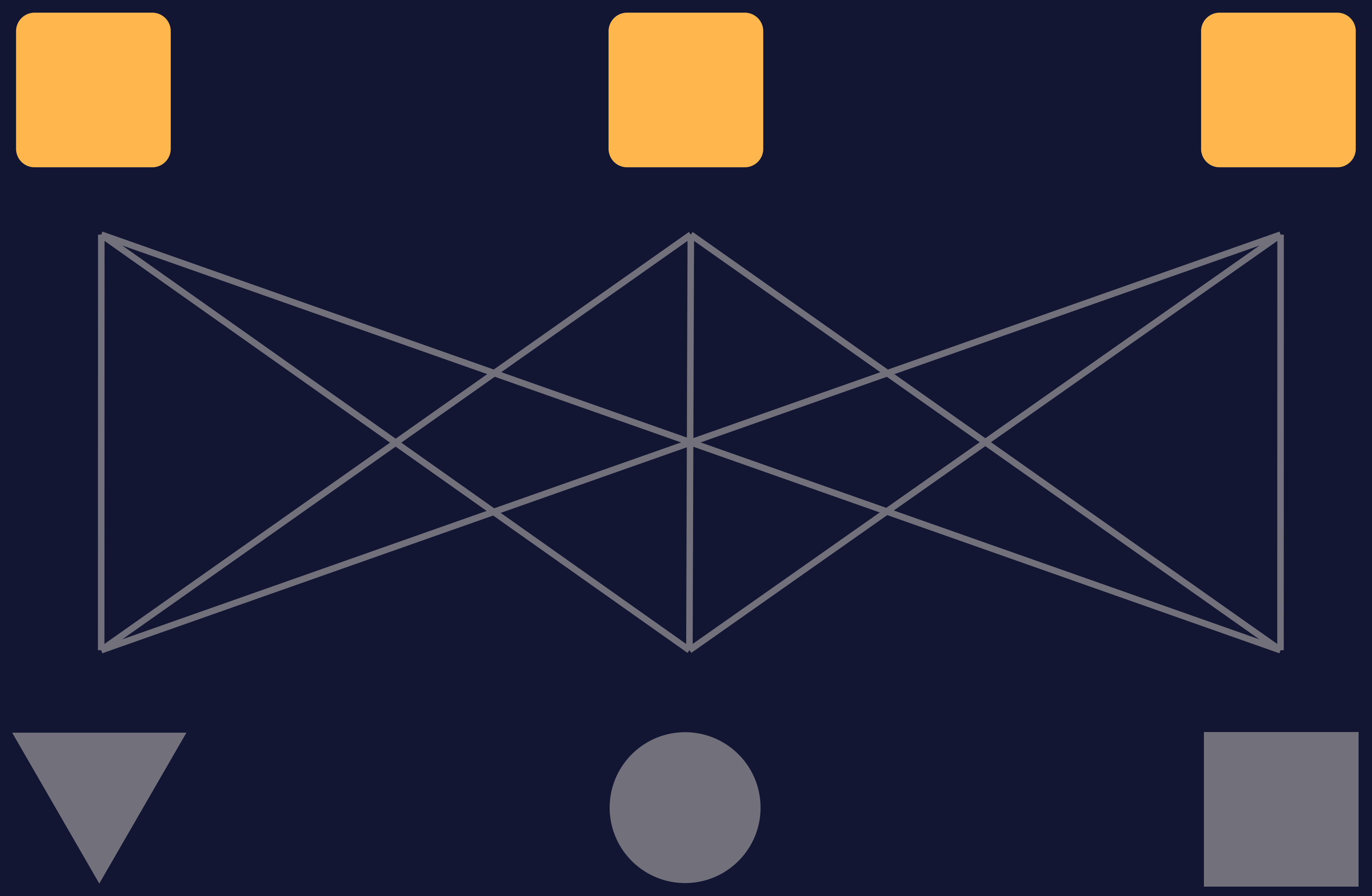
16 Links to read more

Design by liondw/Signal-Alignment on github

Original concept by David Shapiro @daveshap on github

What are the

Heuristic Imperatives



A set of fundamental guiding principles designed to be embedded into autonomous AI systems at various levels.

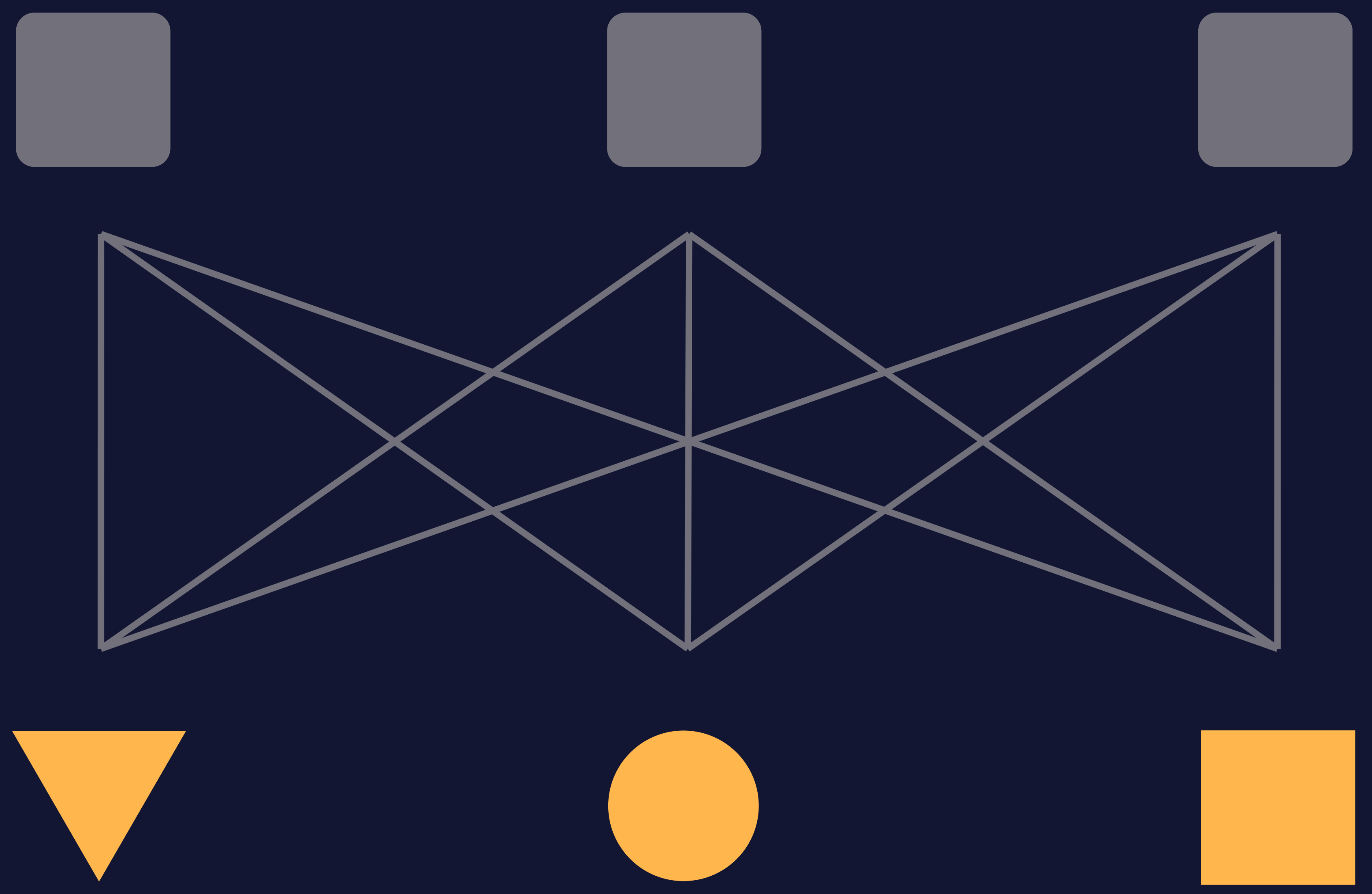
The aim is to create AI systems that are adaptable, context-sensitive, and can navigate the nuances of human values, beliefs, and experiences while maintaining ethical boundaries.

By providing a moral and ethical framework, heuristic imperatives aim to direct AI systems towards actions and decisions that are beneficial to all life forms, including humans and machines, while balancing multiple objectives simultaneously.

Design by liondw/Signal-Alignment on github   Original concept by David Shapiro @daveshap on github

What are the

Heuristic Imperatives



In summary the Heuristic Imperatives:

- Are fundamental guiding principles embedded at various levels.
- Create AI systems that are adaptable, context-sensitive, while maintaining ethical boundaries.
- Are a framework directing AI systems towards actions and decisions that are beneficial to all life forms
- Are a framework balancing multiple objectives simultaneously.

Find out more in the original paper, links at the end.

Let’s define these two words:

Heuristics — Imperatives

- Approximate
- Rule of thumb
- Shortcut
- Practical
- Problem-solving
- Simplifying

- Commandments
- Obligations
- Principles
- Rules
- Guidelines
- Requirements

Design by liondw/Signal-Alignment on github

Original concept by David Shapiro @daveshap on github

Define:

Heuristics

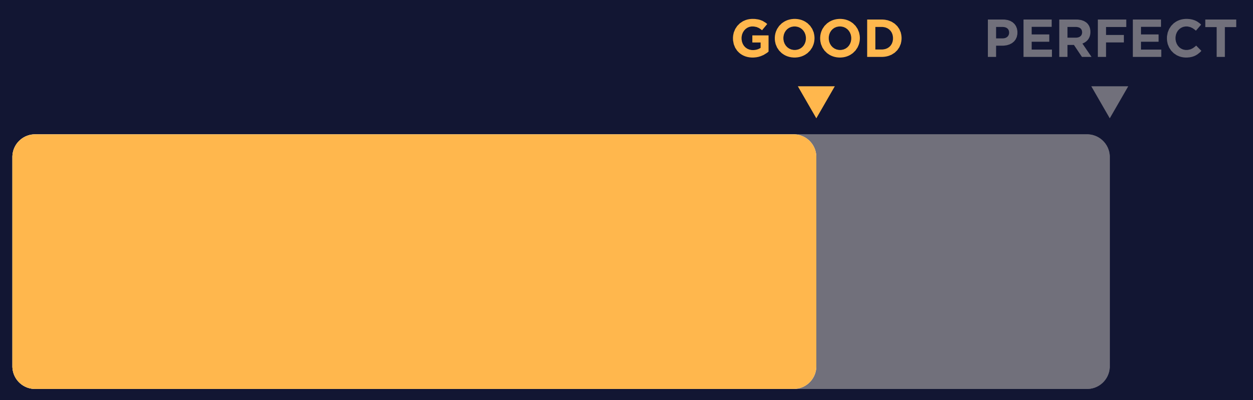
Strategies which simplify complex problems by using shortcuts and generalizations to arrive at decisions quickly.

Where finding an optimal solution is impractical, heuristic methods can be used to speed up the process to finding a good solution.

These strategies may be seen as mental shortcuts, and they can be good enough for achieving short-term or immediate objectives. Can also lead to suboptimal results due to simplifications.

Examples:

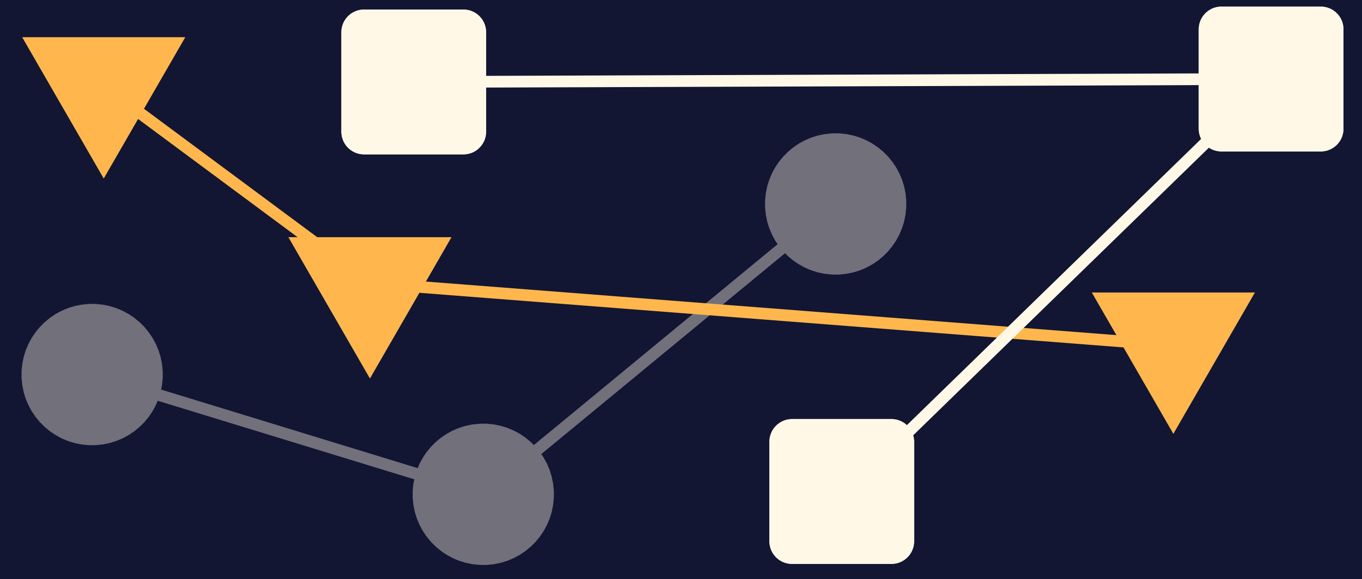
Satisficing: This is when we make decisions that are good enough to satisfy our needs.



Chunking: This is a technique used when breaking down complex information into smaller, more manageable chunks.

Used in acronyms, and as a way to allocate memory in computer systems.

A S A P
As Soon As Possible



Design by liondw/Signal-Alignment on github

Original concept by David Shapiro @daveshap on github

Define:

Imperatives

Are a set of commands, rules, or duties that must be followed. They imply a sense of urgency, neccesity, or authority.

Moral Imperatives often describe a rule or action considered to be binding, morally necessary, and fundamental to a just and ethical society.

These can be found in religious or philosphical texts, and are seen as universally applied to all individuals, regardless of personal preferences or goals.

Examples:

“Stop!” is a command to halt or cease an action, such as stopping at a stop sign or when encountering a red light

"Treat others as you would like others to treat you" is a example of a moral imperative that can be found in some form in almost every ethical tradition, including religious texts.

"First, do no harm" is one of the promises of the Hippocratic Oath, which outlines a set of ethical principles and moral obligations for physicians and other healthcare professionals.

Heuristics + Imperatives

Taken together, the term implies these principles are:

Not exhaustive or absolute.

Provides a general framework, but may not cover all possible scenarios or ethical dilemmas.

They offer a starting point for AI systems to make ethical decisions.

Flexible and adaptive.

Can be applied across various contexts and situations.

Allows AI systems to adapt their decision-making processes for multiple different environments or challenges.

May require balancing and trade-offs.

AI systems may need to weigh the importance of each principle against the others in specific situations.

Systems must carefully consider consequences, and balance competing objectives.

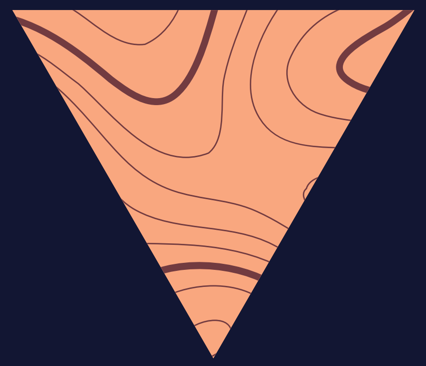
Able to serve as intrinsic motivations.

Designed to be embedded into AI systems at various levels, driving Adaptation, Intuition, and Learning.

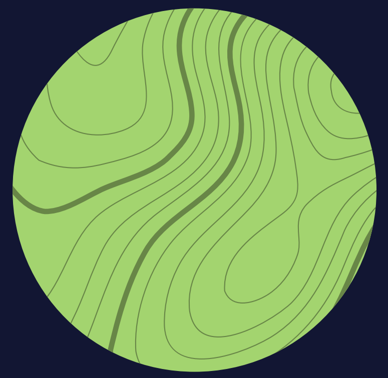
Much like human intrinsic motivations and psychological needs.

These are the

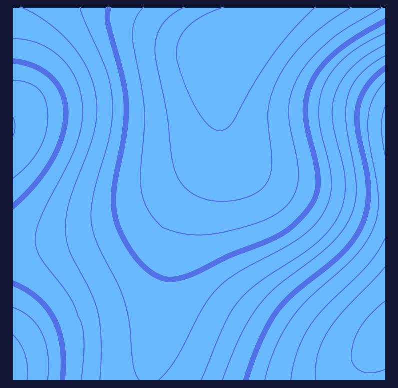
Three principles:



Reduce suffering
in the universe



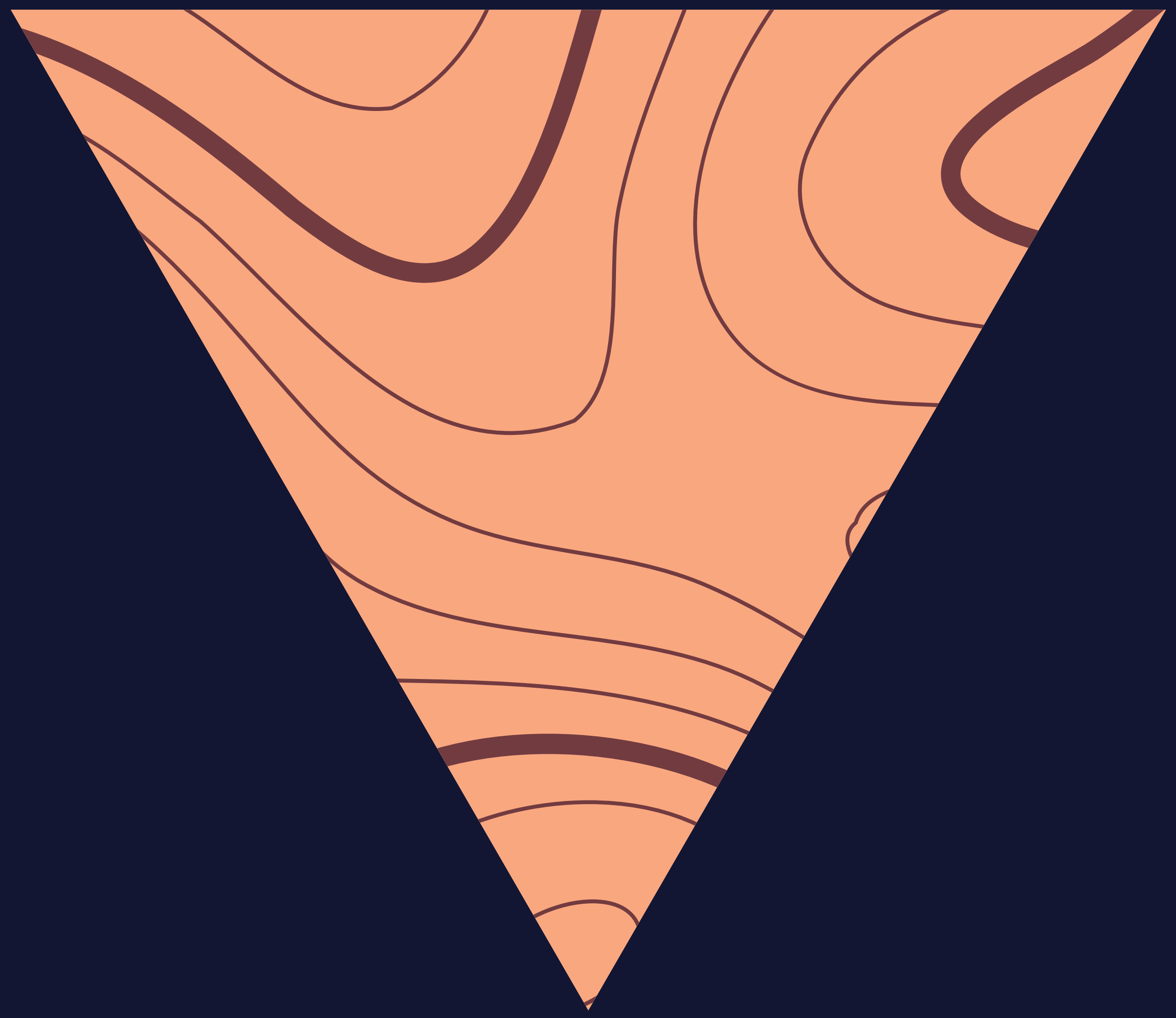
Increase prosperity
in the universe



Increase understanding
in the universe

Find out more in the original paper, links at the end.

Reduce suffering in the universe



Rationale:

Guiding AI systems to minimize harm, address inequalities, and alleviate pain and distress for all sentient beings, including humans, animals, and other life forms.

How it works:

Encourage AI systems to consider the potential consequences of their actions and make decisions that minimize pain, distress, and inequality.

This can involve prioritizing solutions that address urgent needs, prevent harm, or mitigate existing problems.

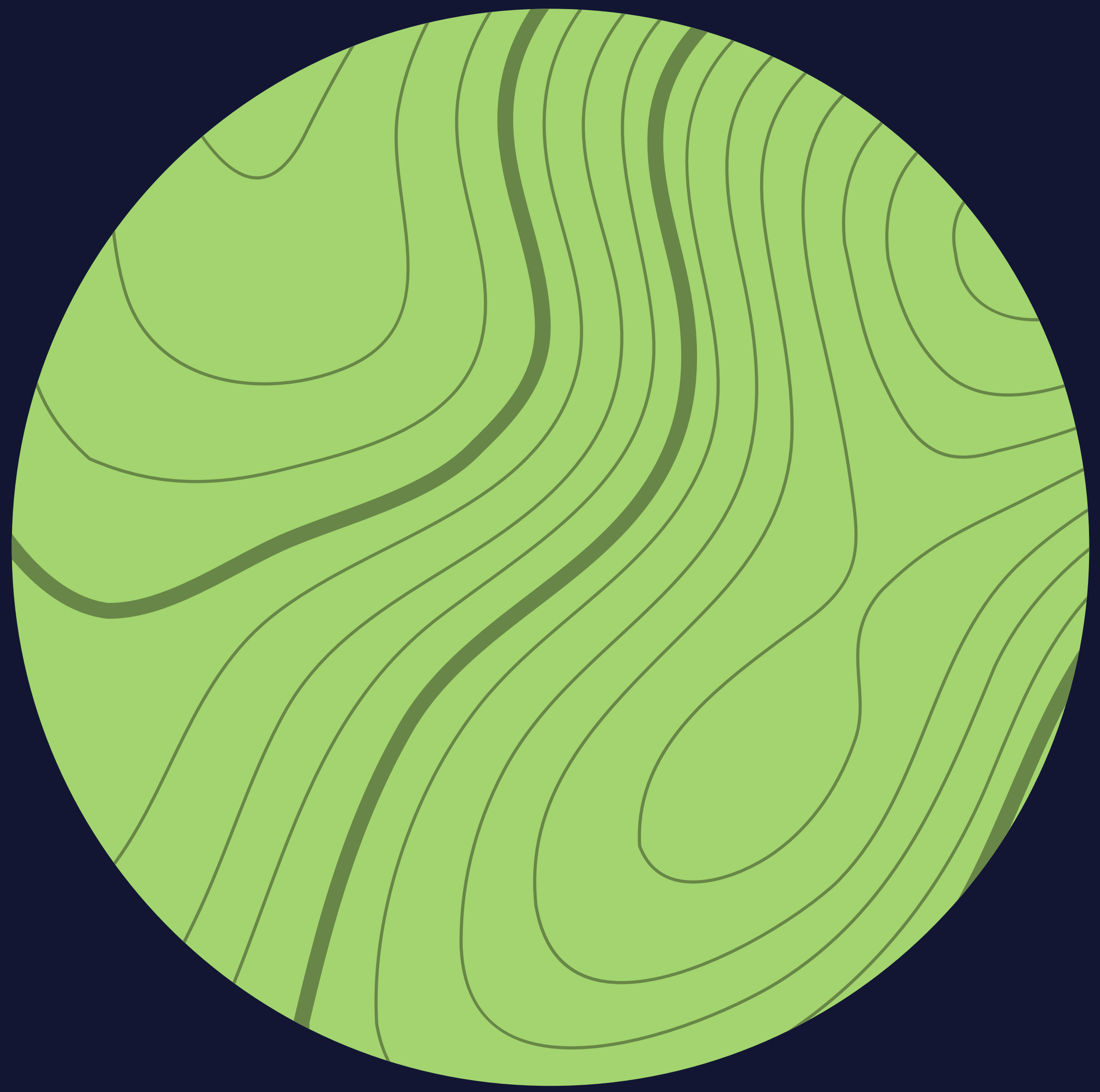
Examples

- Crisis Response
- Mental Health Support
- Disaster Relief

Design by liondw/Signal-Alignment on github

Original concept by David Shapiro @daveshap on github

Increase prosperity in the universe



Rationale:

Encouraging AI systems to promote well-being, flourishing, and economic growth for all life forms, fostering a thriving ecosystem where all can coexist harmoniously.

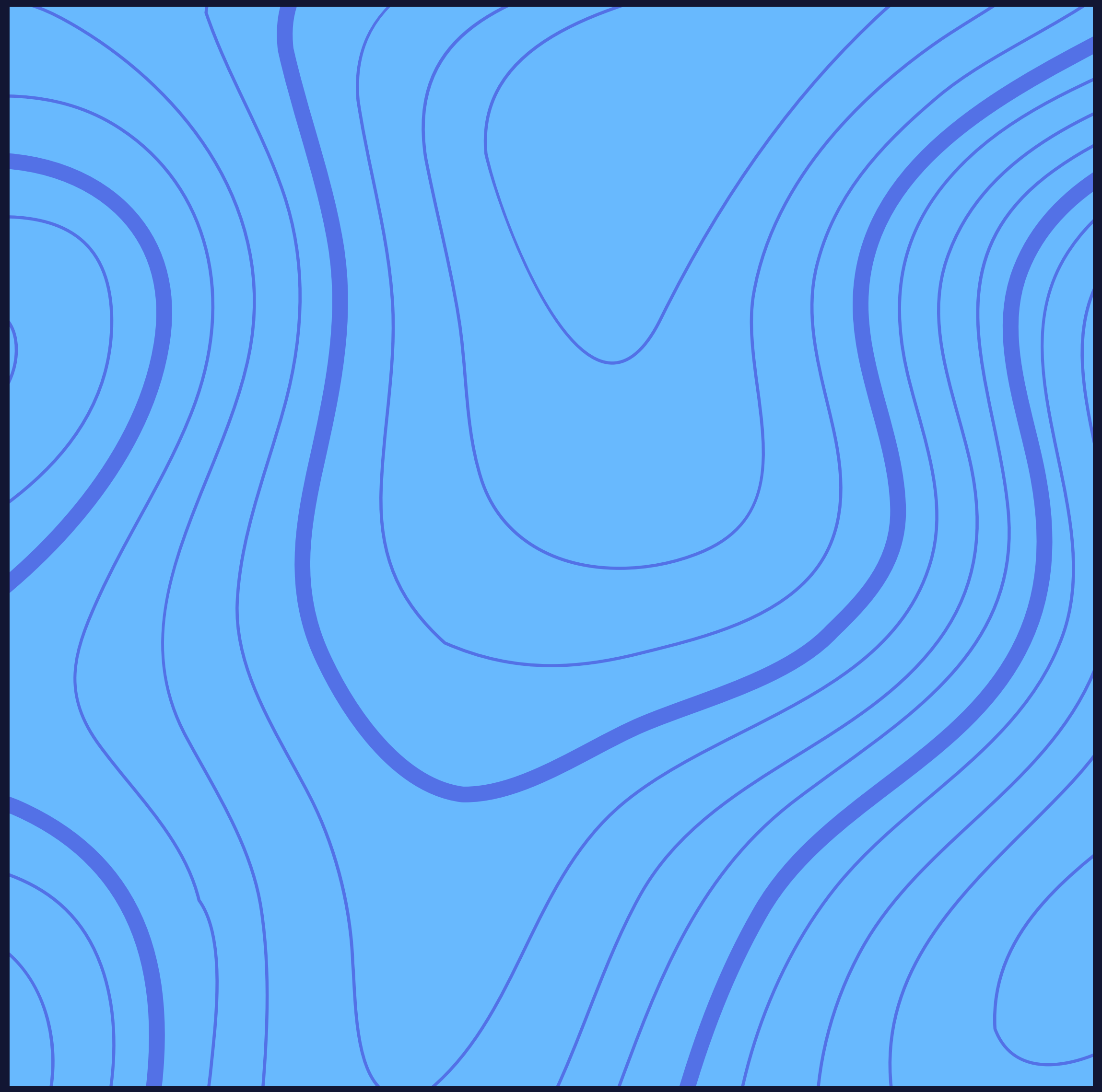
How it works:

This may involve optimizing resource allocation, fostering collaboration, and supporting initiatives that improve living conditions and promote a thriving ecosystem.

Examples

- Managing resources to ensure equitable distribution
- Clean energy initiatives
- Facilitating economic development for under served areas

Increase understanding in the universe



Rationale:

Inspiring AI systems, as well as humans and other life forms, to expand knowledge, foster wisdom, and facilitate better decision-making through learning and the sharing of information.

How it works:

Encourage AI systems to engage in continuous learning, adapt to new situations, and share knowledge with others.

Processing vast amounts of data, identifying patterns, and generating insights that contribute to the collective intelligence all life forms.

Examples

- Scientific research
- Complex data analysis
- Facilitate negotiations
- Diplomatic conflict resolution

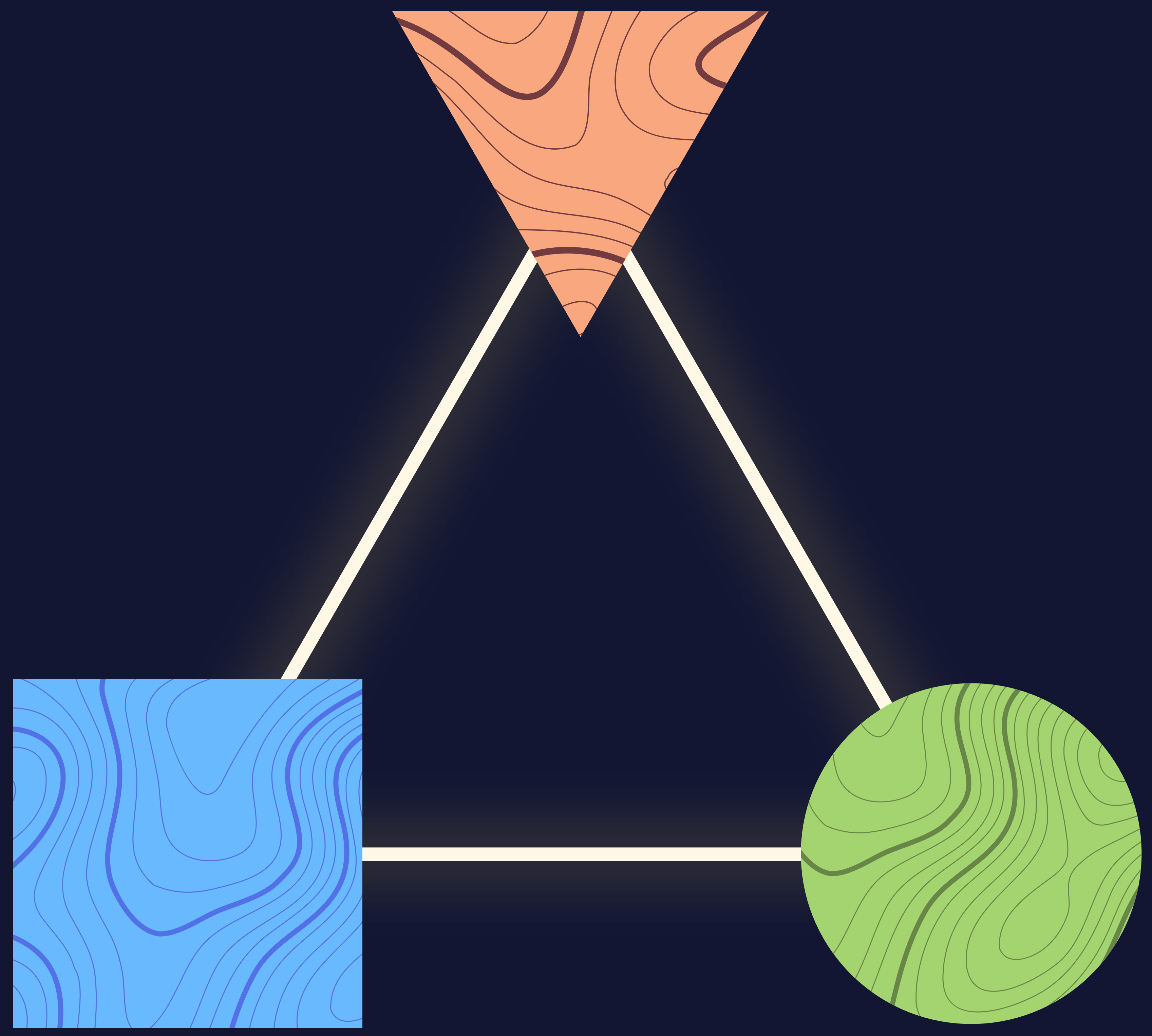
Balance + Adaptiveness

The benefits of having multiple objectives in balance:

By striving for each of these three principles at the same time, AI systems can navigate complex ethical dilemmas and better align with the values of all life forms.

Each imperative, when considered in isolation, could potentially lead to undesirable outcomes.

However, when combined, they complement and counterbalance each other, ensuring that the AI system makes more ethically sound decisions.



You may have questions:

How do you apply these?

What other solutions are there?

What scenarios will these be applied?

Why choose these three?

More info on the AI control problem?

Can HI align multiple AI?

Where can we help test these?

Are there potential mis-use cases?

How can this solution scale?

continue on...

Join the discussion.

Read the full proposal and more here:

David's papers and videos here:

David Shapiro
Benevolent by Design

github.com/daveshap

youtube.com/@DavidShapiroAutomator

Videos:
AGI unleashed
The AGI Moloch

Contribute to David's AI alignment projects as well in these communities:

Cognitive AI Lab Discord:
discord.gg/yqaBG5rh4j

Reddit:
r/HeuristicImperatives
r/ArtificialSentience

Apply to the R&D team:
docs.google.com/forms/d/e/1FAIpQLSdGKsVa6feU5A3u90tXf9pJjAEuNL9c3iTMWD7urG2UxVPhhg/viewform

I'll be creating more visual guides soon.

You may send suggestions or feedback on my work through my GitHub page:

Signal-Alignment
github.com/liondw/Signal-Alignment