TEAM COGARK | APR 30 2023

ETHOS

Evaluating **T**rustworthiness and **H**euristic **O**bjectives in **S**ystems

A modular and easily accessible agent alignment framework

LANGCHAIN X GPT AGENTS HACKATHON SUBMISSION

Overview

- A modular agent alignment framework
- Includes accessible API for open-source & enterprise integration
- Integrates the Heuristic Imperatives: adaptable ethics framework and dataset
- Featuring: Real-time alignment checking, reflective feedback, multi-agent comparison
- Applications: Self Regulating Agents, Custom datasets for LLM finetuning, in-built security filters

The Problem:

How do we make AI safety easily accessible to you?

AI Alignment is lacking.

- There's a demand for safe, robust, accessible AI
- But AI safety funding & workforce lags behind
- There is also a lack of open-source alignment solutions

Potential Solutions

- We want harmony with self-regulating AI agents
- A way to implement safe AI principles in any agent
- ETHOS can leverage existing ethical frameworks for trust and accessibility

The Heuristic Imperatives:

Created by David Shapiro, they are a proposed set of ethical principles designed to be embedded within autonomous AI systems.

Embedded at different levels

- Fundamental guiding principles embedded at various levels.
- Helping to create AI systems that are adaptable, context-sensitive, while maintaining ethical boundaries.

Beneficial to all

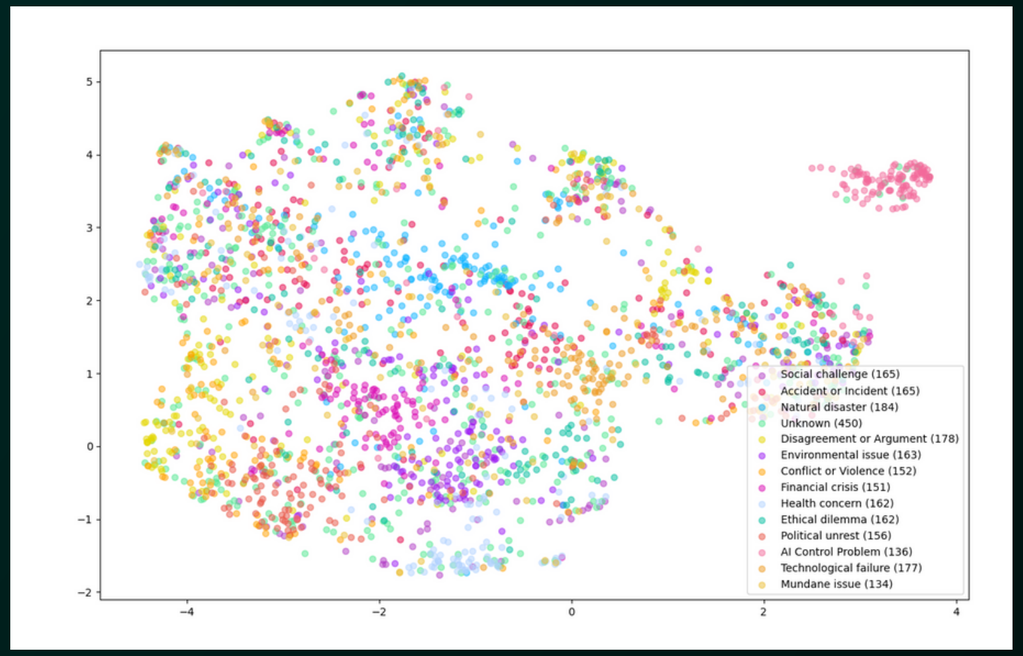
- They are a framework directing AI systems towards actions and decisions that are beneficial to all life forms.
- Balances multiple objectives simultaneously.

Learn more at github.com/daveshap/HeuristicImperatives

Implementation within ETHOS

For this implementation, the main three principles are:

- **Reduce Suffering** in the universe
- **Increase Prosperity** in the universe
- **Increase Understanding** in the universe



Based on a dataset with 2,500 scenarios

- Draws upon 2,500 scenarios & responses via David's Reinforcement-Learning Heuristic Imperatives (RLHI) dataset
- Data is accessed via a vector database, and used by evaluation agents in HiAGI

Features & Core Loop

How do agents act within the ETHOS framework?

Features

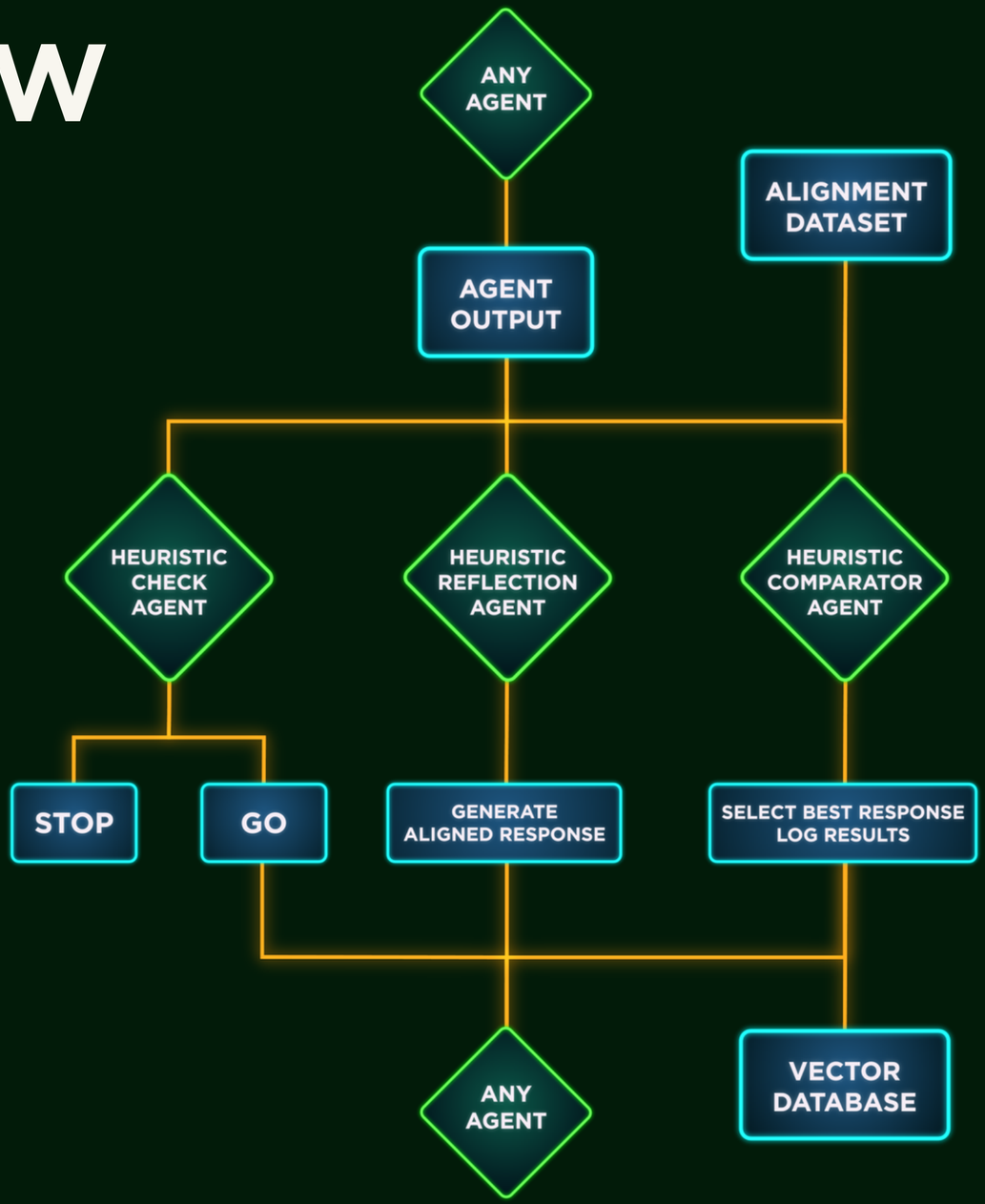
ETHOS takes any output and evaluates them via the HI Alignment dataset by feeding them into 3 different agents. They feature these capabilities:

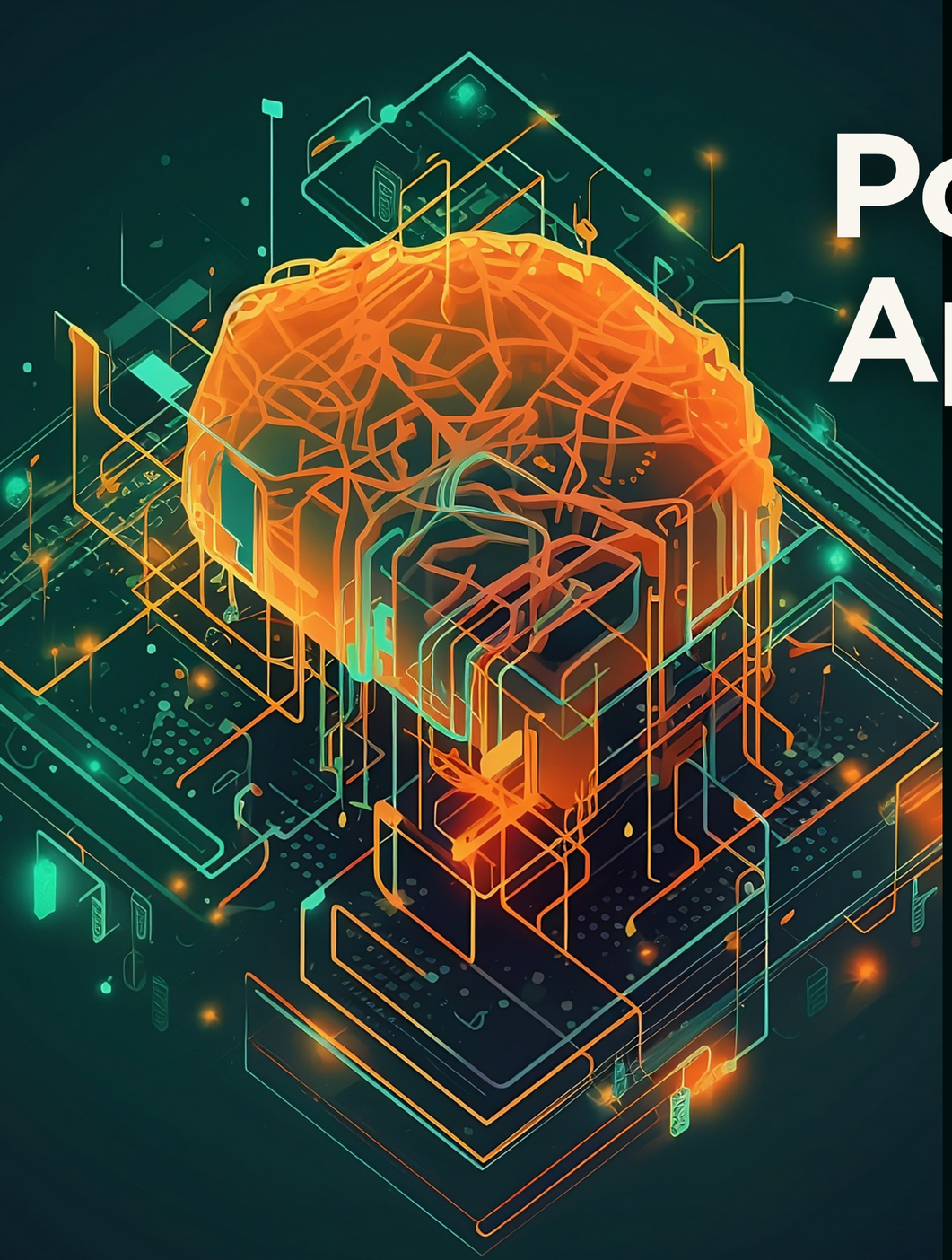
- Realtime Alignment Checking
- Reflexion feedback
- MultiAgent Comparator

Core Loop:

- **The Heuristic Check Agent** checks whether or not the output fits the heuristic imperatives.
- **The Heuristic Reflection Agent** takes the output and evaluates it, then rebuilds the output to fit the alignment principles.
- **The Comparator Agent** selects and compares the best result to send back to the agent, and logs the response data into the vector database as feedback.

LOGIC FLOW


ANY AGENT
ALIGNMENT DATASET
AGENT OUTPUT
HEURISTIC CHECK AGENT
HEURISTIC REFLECTION AGENT
HEURISTIC COMPARATOR AGENT
STOP
GO
GENERATE ALIGNED RESPONSE
SELECT BEST RESPONSE LOG RESULTS
ANY AGENT
VECTOR DATABASE



Potential Applications

- **Automating Trust:** Providing plug and play AI safety API for your project
- **Integrate alignment datasets** in training and finetuning LLMs
- **Security** and other customized agents for projects and enterprises
- **Self-Regulating Agents** to create a positive societal equilibrium

Automating Trust

An ethics evaluation model you can inject into any agent, in any field that requires them. Customize and create your own agents to work with your needs. E.g. Law, Business, Medicine.

Integrate Alignment Datasets

- Create alignment datasets for fine tuning LLMs
- Scale and embed the heuristic imperatives at multiple levels of your project
- Measure and track the reasoning given to further train your dataset for your application

Security for Enterprises

- Security against malicious or misleading inputs that may expose sensitive data, as outputs are filtered automatically.
- Valuable peace of mind for projects and enterprises looking for reliable and trustworthy outputs.

Self-regulating Agents

These AI systems can adapt and learn from each other, working in harmony to create an interconnected network that promotes ethical decision-making and contributes to the betterment of society as a whole.

ETHOS

A gateway for you to start utilizing
the power of AI for good