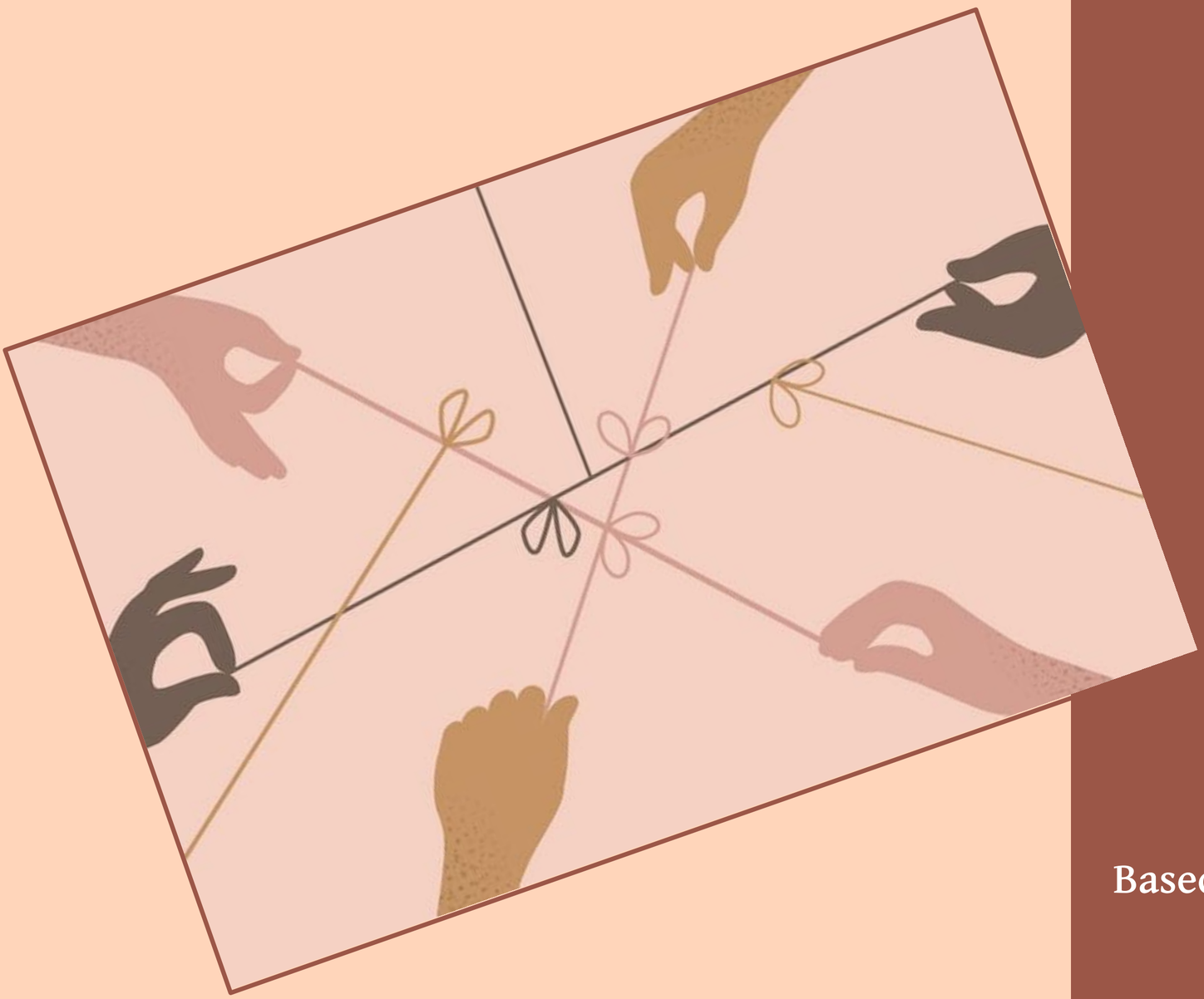

BIRDS x Safety & Alignment bring you

AI Alignment Cohort

Based on the ARENA Curriculum by Callum McDougall

Meet the Team

Akanksha

Discord: akankshanc
Twitter: akankshanc

Herumb

Discord: krypticmouse
Twitter: krypticmouse

Caroline

Discord: c.s1693
Twitter: Shamiso28163594

Sunitha

Discord: prisca6117
Twitter: @sunitha_selvan

Alif

Discord: biggmon
Twitter: @alifmunim

Cohort Contents

Module 1

Basics of Math and Neural Networks
Ray Tracing
CNNs and ResNets
Optimization and Backpropagation, Autograd
GANs and VAEs
EINOPS Library

Module 2

Transformers from scratch
Introduction to Mechanistic Interpretability
Superposition and Sparse Autoencoders
Function Vectors and Model Steering

Module 3

Introduction to Reinforcement Learning
Q-Learning and DQN
Proximal Policy Optimization
Reinforcement Learning from Human Feedback

Module 4

Capstone Project

Participants will make a Group of 3 and work together on a project

Cohort Structure

2 Weekly Sessions

We will give you Pre-Reads

You will give us assignments

Session 1

MONDAY
11 am
Pacific Time

Session 2

Friday
10 am
Pacific Time

Week 1

Monday 7/15: Math Foundations

Friday 7/19: Math Foundations

Week 2

Monday 7/22: Neural Network Concepts & EINOPS

Friday 7/26: Ray Tracing, CNN and ResNet

Week 3

Monday 7/29: Optimization, Backprop, Autograd

Friday 8/2: GANs and VAEs

Week 4

Monday 8/5: Transformers from scratch

Friday 8/9: Introduction to Mechanistic Interpretability

How will you benefit the most from this Cohort?

Assignments
+
Project

Weekly assignments
to be submitted
via
Google Forms
and
Colab notebooks

qu*AI*dditch

Are you the seeker?

Answer maximum
questions during the
weekly session to win the
Golden Snitch
award

Any Questions?