Arun Jose

✉ jozdien@gmail.com • Trivandrum, India • ☎ (+91) 8075727965 • jozdien.com • in Jozdien • github.com/Jozdien

Education

Computer Science and Engineering - B.Tech
College of Engineering Trivandrum
08/18 - 06/22 • CGPA: 8.65

Indian School Certificate
Loyola School Thiruvananthapuram
06/05 - 03/18 • ISC: 91.6%

Awards

ELK Contest
Wrote a prize-winning strategy
"Train a sequence of reporters"

EVOKE'19 Hackathon
Winner, out of 15+ teams
National level tech summit by IEDC & IEEE SB TKMCE

Reboot Kerala Hackathon
2nd Runner Up, out of 30+ teams
State level Hackathon series by the State Department of Higher Education

Pass the Code
Winner, out of 15+ teams
State level tag-team competitive coding competition by IEEE SB GECBH, IEDC, and ISTE

AKCSSC Web Design Contest
1st Place
State level web design competition by IEEE CSKS

Technical Skills

Experienced
Python, React + Native, JavaScript, HTML, CSS

Comfortable
Tensorflow, PyTorch, Java, NextJS, Firebase

Acquainted
C, C++, Assembly (x86), Octave

Experience

09/2022 - Present
Independent Alignment Researcher
- I'm working on technical projects related to interpretability, inner alignment, and related ideas, researching simulator theory, and learning more about current interpretability work.

08/2021 - 08/2022
ML Research Intern
Median Group
- Worked on conceptualizing a tool using GPT-3 for providing real-time conversational analysis as an external moderator with a lean against structuralism, posturing, and the like, keeping them rich in object-level content.
- Created an ML pipeline for analyzing and classifying segments from audio conversations into clusters based on their argumentative quality and structure.

01/2022 - 06/2022
AI Safety Camp
- Worked on Kyle and Laria's team to model alignment failures in GPT-3 and explore aspects of simulator theory.
- Primarily did conceptual work on gradient filtering, as well as testing GPT's ability to model increasingly complex values and recognize dangerous power-seeking strategies.

06/2022 - 07/2022
ML Alignment Theory Scholars - Deceptive AI
Stanford Existential Risks Initiative
- Took part in the training program and research sprint, mentored by Evan Hubinger.
- Wrote this post evaluating the conceptual and practical potential of generative models to accelerate alignment research.

Projects / Publications

Isolating Updates (Stable Baselines3)
Ongoing interpretability project on isolating an update signal corresponding to the high-level objective structure in a deep RL network.

Reward side-channels *(SpinningUp)*
Ongoing explorative work into empirical analysis of internal reward representation in reinforcement learning agents to better understand the mechanics of deceptive alignment. I work on project management as well, and oversee another developer working on this.

Metise *(Preprint; Tensorflow)*
Reversible colour-density image compression using conditional adversarial networks.

Calibration Game *(React Native)*
Android app operating under the rules of the credence calibration game. I handled design, development, and deployment.

Aumann's Game *(NextJS, Node)*
A multi-player web version of the calibration game. I handle design and code the front-end.

Veris *(NextJS, React Native, Node)*
Open-source web/app framework for college administration, deployed in CET.

Karma *(HTML, PHP)*
System to facilitate fast rescue during emergencies. 1st Place at national-level EVOKE'19 Hackathon.