[Confidential - Do not distribute]

3-year Research Roadmap

Center for Ethical Intelligence (CEI)
C. Jimmy Lin, MD, PhD, MHS, MAR
11 November, 2025

Goals & Objectives

Genesis 1:26-28 Then God said, “Let us make man in our image, after our likeness. And let them have dominion over the fish of the sea and over the birds of the heavens and over the livestock and over all the earth and over every creeping thing that creeps on the earth.” So God created man in his own image, in the image of God he created him; male and female he created them. And God blessed them. And God said to them, “Be fruitful and multiply and fill the earth and subdue it, and have dominion over the fish of the sea and over the birds of the heavens and over every living thing that moves on the earth.”

ACADEMIC & SCHOLARSHIP

Field of AI

- Create **database** of existing benchmarks, evals, datasets, and corpora, and organize and review existing research
- Create **new benchmarks** for values alignment that can be used for measurement and assessment - ultimately for adoption for SOTA models
- Produce large **data sets** that can be used for pre-training, fine-tuning, RL, or other steps for model development
- Produce additional **scholarship** and papers at AI conference (e.g. ICML, NeurIPS, ICLR) to add to work in the field of values alignment

TALENT PIPELINE

People

- Foster a network of expert scholars in values alignment research - goal of top 1% of researchers given knowledge and reps
- Create and teach courses at different academic centers and online about values alignment
- Create internships for entry-level AI scientists to contribute
- Create scholarships and awards to fund AI alignment research projects

SPIRITUAL & FAITH

Church

- Fellowship of Christians working in scholarship of AI/ML that can support and pray for each other, and be discerning to God’s work in AI
- Ways for the Christian church to be active in shaping the field of AI - not just in principle statement, but create actual data that can be used
- Contribute as public intellectuals to help the church not to fear and ignore AI, but see how God can use AI for his purposes - Redemptive AI

Research Roadmap: Monthly

ORGANIZATIONAL | MODEL CARDS, RESEARCH REPORTS | LITERATURE: BENCHMARKS & EVALS | DATA GATHERING & SCHEMA

 Oct

- Finalize and **gather** interested researchers
- Kick off weekly **research meeting** and logistics
- Team get to know each other
- Provide view of 3-year **research roadmap** to team
- Determine structure and composition of **research pods**

 Nov

- Explore documentation of **frontier models** from AI labs including Google Deepmind, OpenAI, Claude, Chinese models (+ history)
- Focus on approaches to **value alignment**
- Understand the papers for **utilized benchmarks** and evals

 Dec

- For each of the approaches: **Ethics** (virtue, deontological, consequentialist), **Moral Psychology** (MFT, etc), **Cultural** (Country, ethnic groups), **Religious** (Christian, Islamic, Buddhist, etc)
- Comprehensive **literature review** the approach to value alignment within last 5 year

 Jan

- Compile the **data sets and schema** used in the literature
- Incorporate and **critique** existing methods and dataset
- Make a **proposal** for more accurate and in depth data sets
- Identify possible **experts** or approaches to contribute

DELIVERABLES / OUTPUTs

- Scholar profiles
- 3 year research roadmap
- Research pod formation
- Background resources on Alignment

- Overview of frontier models approach to value alignment
- Understanding of dominant benchmarks and evals and approaches to alignment

- Comprehensive literature review of a subfield and list of associated benchmarks, evals, datasets, corpora for overall database

- Combined data set for specific sub-field
- Approach to create gold standard datasets with associated experts

First Year Research Roadmap: Quarterly

BACKGROUND, LITERATURE

Q1

- Level up group on **background** of alignment, model development
- Overview of approach to alignment from the **top AI labs**, focusing on value alignment
- Deeper dive on the **literature** of common benchmarks and evals
- Deeper dive on different **subfields** including ethics, moral psychology, religious, and cultural
- Literature **review**, data set **compilation**, **schema** building, and **expert** identification

DATA GENERATION, DATABASE

Q2

- Creation of **overall database** of benchmarks, evals, datasets, and corpora for moral reasoning and values alignment
- Draft **review paper** of all different types of values alignment (example) with associated database, each pod writes a chapter, each scientist writes a section
- Creation of **new datasets** for each of the subfields - outreach, data generation

EXECUTION & COMPARISON

Q3

- Execute **historical benchmarks** of relevant LLM - historical and SOTA
- Apply the **newly designed** benchmarks on the same set of models
- Data science project: **comparison** of different benchmarks and evals
- Drafting of **abstracts** and initial findings for possible submission to different AI conferences

FINALIZATION OF RESEARCH

Q4

- Publication of the **overall review paper**
- Publication of the **sub-field papers** of new benchmarks
- Compilation of the results in **combined database** with Hugging Face hosting
- Determination of how the different approaches can be combined - **overall schema**
- Determine **best practices** to expansion of other datasets

3 year Research Roadmap: Annual

FOUNDATION

Y1

- Comprehensive understanding of **existing AI** lab use of alignment, focusing on values alignment
- Comprehensive deep dive of **sub-fields of values alignment** including ethics, moral psychology, religious, and cultural
- **Literature** review of existing work and evaluates on SOTA models and compile **database** of existing benchmarks, evals, datasets, and corpora
- Creation of **new schema, datasets, benchmarks** and execute on SOTA models start drafting **abstracts** for submission to conferences

APPLICATION

Y2

- **Publish**: 1) overall review paper, and 2) sub-field specific deep dive papers and new benchmarks, evals, and datasets
- Depending on team size, perform **second round** of sub-field exploration
- Start design on **faith-based ethical frameworks** and start engaging with faith communities with data collection and benchmark building
- **Assess** new benchmarks and datasets of **SOTA models** and **publish** second batch of papers to conferences
- Establish **database** of values alignment to be central resource for the field

SCALE

Y3

- **Publish**: 1) second batch of sub-field exploration, 2) initial faith-based benchmarks of SOTA models
- Implement large-scale **database** alignment benchmark research, enable automated assessment of benchmarks on platforms such as Hugging Face
- Engage **multiple faith communities** for large-scale data collection on path to create definite datasets and benchmarks for the community
- Possible **expansion** ideas: moral reasoning, multiturn, multimodal, agentic, language, combination of different sub-fields

Ways to be Involved

Col 1:9b-12 We have not ceased to pray for you, asking that you may be filled with the knowledge of his will in all spiritual wisdom and understanding, so as to walk in a manner worthy of the Lord, fully pleasing to him: bearing fruit in every good work and increasing in the knowledge of God; being strengthened with all power, according to his glorious might, for all endurance and patience with joy; giving thanks to the Father, who has qualified you to share in the inheritance of the saints in light.

RESEARCH PODS

SCIENTIST / ENG.

- Time commitment: 1-3 hr / week
- Responsibility: Help execute research program with supervision
- Experience: MS or PhD in training, few years of industry experience, willingness to grow and learn in new area
- Publication Authorship

RESEARCH FELLOW

- Time commitment: 2-4 hr / week (weekly group meeting, sub-mtg)
- Responsibility: Contribute to research group, own academic contribution
- Experience: PhD+, or 3-5 years of industry experience, published papers
- Publication Authorship

TECH/TEAM LEAD

- Time commitment: 3-5 hr/week (weekly group meeting, run own mtg)
- Responsibility: Lead a small research group, set agenda for group
- Experience: leading research groups, timelines, and publications
- Publication: First or senior position

Academic Growth: Deep team learning in specific subfield, deep technical expertise of alignment, growth to top 1% of field
Spiritual formation: Consistent exploration of integration of faith and learning, fellowship and community growth

CONTRIBUTOR

- Time commitment: Ad hoc
- Responsibility: Specific tasks as needed
- Experience: Relevant expertise e.g. software engineering, ethics, theology

ADVISOR

- Time commitment: 1-2 hr/month
- Responsibility: Provide advice and feedback on research progress
- Experience: Significant experience in

OTHER?

- We are still building the program and are open to different ideas on involvement and contribution

Growth: Awareness and keeping up to date of whole field, fellowship and access to community resources

STAYING CONNECTED

- **Whatsapp**: More informal chats, latest research news, quick updates, news, celebrations, prayer
- **Email Listserv**: Longer updates & announcements

Will this work? How to increase probability of success?

Psalm 127:1 Unless the Lord builds the house, those who build it labor in vain. Unless the Lord watches over the city, the watchman stays awake in vain.

QUESTIONS	RESPONSES
⚠ Risks	Mitigations
As startup **organization**, how do we know it will succeed?	• Jimmy is a serial social entrepreneur and has built a volunteer non-profit Rare Genomics Institute, which operates at its peak 200 volunteer staff across 7 countries. We have helped hundreds of rare disease families, funded millions of dollars of research, have active programs in AI and bioinformatics. • Dawn is a serial entrepreneur and has build an edtech hardware startup and a content media business.
Can a group of **volunteer** researchers produce results?	• At Rare Genomics, we have published 10+ papers - with PhD experts with help from interns and non-technical staff - even including a 2 books on rare diseases.
How do we make sure team is appropriate **staffed**?	• Up front, important to select talent with the right expertise and availability, setting expectations • Not everyone turns out to be good fit, make sure we have buffer for over-staffing
How do we appropriately **scope** and **design** the projects?	• Make sure that the project can scale with appropriate level of effort - modular • Start with literature review - lower risk, and then build to new efforts
How do we know that the work will be at **quality** for AI conferences?	• Team leads will have experience in publishing and leading efforts • Advisors can help ensure the level of rigor of the work
How do we know that the work will have **impact** at the AI labs?	• We are in active conversation and collaboration with leaders in the AI labs • Team members have experience to have benchmarks incorporated into SOTA model evals

Meeting Rules

PURPOSE & POSTURE

- **Remember why we're here**: We gather to pursue truth, advance knowledge in AI, and honor God in how we think, speak, and act together.
- **Assume good intent**: We start from the assumption that others are sincere, even when we disagree with their ideas or conclusions.
- **Respect each person's dignity**: Every participant is made in the image of God and must be treated with courtesy, patience, and kindness—regardless of rank, background, or viewpoint.

CONFIDENTIALITY & TRUST

- **No sharing of information outside meeting**: Do not share information (about research subjects, students, colleagues, institutions, or partner organizations) without explicit permission
- **No recording without consent:** Audio/video recording, screenshots, or transcript sharing require prior, explicit consent from everyone present
- **Ask before sharing**: Ask for permission before sharing any content from meetings

BRINGING OUR WHOLE SELVES

- **Presence**: Please be fully present at the meeting. Try not to multitask. If possible, please turn on your camera
- **Make space for every voice**: Senior scholars and frequent talkers intentionally leave room for quieter members, junior researchers, and those from underrepresented backgrounds.
- **Critique ideas, not people**: Disagreement is welcome; disrespect is not. Phrase pushback as engagement with an argument, not a judgment of someone's character, intelligence, or faith.

MEETING LOGISTICS & RESPONSIBILITY

- **Start and end on time**: We honor each other by keeping to agreed start and end times, and by sticking to the agenda as much as reasonably possible.
- **Come prepared**: If work is to be presented to the group, please make a good-faith effort to deliver what was committed
- **Follow-up is explicit**: Action items, responsibilities, and next steps are clearly stated at the end, so people know what's expected of them before the next meeting.

Research Teams

1 Peter 4:8-12 Above all, keep loving one another earnestly, since love covers a multitude of sins. Show hospitality to one another without grumbling. As each has received a gift, use it to serve one another, as good stewards of God's varied grace: whoever speaks, as one who speaks oracles of God; whoever serves, as one who serves by the strength that God supplies—in order that in everything God may be glorified through Jesus Christ. To him belong glory and dominion forever and ever. Amen.

TEAM COMPOSITION CONSIDERATION: Experience, existing relationships, complementary skill sets, availability, time zones

- Will revisit team composition as needed, but also at the beginning of 2026

TEAM 1

NameTBD

Team Lead
- Rui Guo

Research Fellow
- Toby Li

Research Scientist
- Qiuyue "Joy" Zhong
- Xin Chen

Research Advisor
- Mike Lee

TEAM 2

NameTBD

Team Co-Leads:
- Marcus Schwarting
- Andrew Wagenmaker

Research Fellow
- Millie McKaylin
- Ziwen Han*

Research Scientist
- Erik Nordby

Research Advisor
- John Slattery

TEAM 3

NameTBD

Team Co-Leads:
- Nathan Gaw
- Richard Zhang

Research Fellow
- Alex Chao

Research Scientist
- Anji Zhang

Research Advisor
- Emily Wenger

Next Steps: Determine name, communication method, subgroup time, how to divide the tasks, how you want to leverage advisor, Google minutes, etc

* Available mostly on weekends // Link to Scholar Profiles

Research Sprint 1: Understanding the field

Luke 14:28-32: For which of you, desiring to build a tower, does not first sit down and count the cost, whether he has enough to complete it? Otherwise, when he has laid a foundation and is not able to finish, all who see it begin to mock him, saying, 'This man began to build and was not able to finish.' Or what king, going out to encounter another king in war, will not sit down first and deliberate whether he is able with ten thousand to meet him who comes against him with twenty thousand? And if not, while the other is yet a great way off, he sends a delegation and asks for terms of peace.

TEAM 1

NameTBD

US/European
- Anthropic
- Mistral

Chinese
- Deepseek
- Baidu

TEAM 2

NameTBD

US
- Google Deepmind
- xAI

Chinese
- Moonshot
- z.ai

TEAM 3

NameTBD

US
- OpenAI
- Meta

Chinese
- Alibaba
- Tencent? Meituan?

Primary objective: Dive deeply and understand the alignment strategy and specifically values alignment for each of the top AI labs with state-of-the-art large language models. Assessment would include current as well as historical models to show evolution. Sources include model cards, white papers, pre-prints, external assessments, and other scholarly publications. Final output would be a comprehensive analysis paper of alignment at the top AI labs including text component and data science heatmap comparison. Completion goal is draft by end of year (2 months)

Secondary objective: Dive deeply and understand the key papers of benchmarks, datasets, and approaches to values alignment. What are the areas and approaches that are well covered? What characteristics make these datasets to be utilized by these companies? What are the gaps and opportunities? What are the technical alignment approaches used? How can this inform how we create datasets and benchmarks?

Research: Sources & Methods*

SCHOLARLY LITERATURE

1°

- Publication **directly** from the **AI lab**: whitepaper, model cards, arXiv, blog post
- Assessment performed by **external sources** - when the benchmark is introduced, analyzed after the launch
- **Literature review** (topical) that performs the analysis of the different LLMs
- **Tools**: Google Scholar, Semantic Scholar, **Zotero** Cloud (& other reference managers) SAMPLE

DEEP RESEARCH

DR

- **Deep research** tools using LLM models from AI labs
- **Meta-prompting**: "Please write an LLM prompt that deeply researches..."
- **Agentic** research from tools such as Manus AI, Kimi K2, Deepseek R1
- **Tools**: Deep research from **Gemini**, **Claude**, **ChatGPT**, Perplexity, StormAI SAMPLE

AI-ASSISTED RESEARCH

AI

- **Literature research** tools: including Consensus, Elicit, SciSpace, Scite.ai, Iris.ai, Anaea
- **Literature mapping**: Research Rabbit, Litmaps, Connected Papers, Inciteful
- **Writing tools**: Jenni, Paperpal, Writeful, Trinka, Thesify, Avidnote, Paperguide

MULTI-MODAL TOOLS

A/V

- Tools to enhance and process research outputs and research papers
- **NotebookLM** - can create **podcast** overview of Deep Research outputs or specific papers, create consolidated overview from a list of papers, also videos SAMPLE
- **Gamma**: summarize and structure extensive outputs into easy to navigate slide decks
- Speechify: **Text-to-speech** for summaries on the go

***** term used to describe the practice of intelligence collection and analysis, include tradecraft (DOI:10.1093/oxfordhb/9780195375886.003.0004)

SAMPLE Already pre-generated samples for handful of companies, Jimmy will share after the meeting

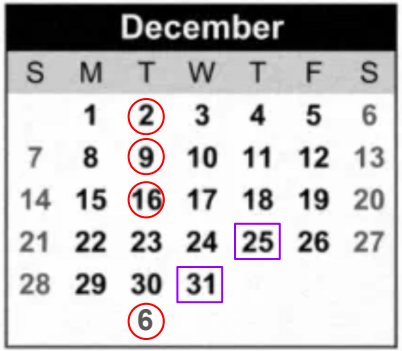
Upcoming Planning & Calendar

GROUP MEETING LOGISTICS

- **Meeting format**: Each research team present highlights and updates in slide format, goal is have more detail on slides than able to present (instead of bullet points on slides and expanded with presentation)
- **Timing**: Three groups to present 20 minutes each
- **Preparation**: Each team meets outside of weekly research meeting and team lead coordinate distribution of work and responsibilities
- **Advisors** to provide feedback offline

PAPER LOGISTICS

- **Writing**: Each member writes 3-5 pages (total around 36-60 pages), team leads supervises and revises their sections
- **Heatmap/Table**: select members to work on comparison, overall views
- **Goal**: First draft to be ready by Jan 6th, 2026



GROUP MEETING TIMELINE

- ➢ Meeting 1: **Kickoff meeting!** Get to know each other
- ➢ Meeting 2: **3 Year Roadmap**, options to participate
- ➢ Meeting 3: **TODAY**! **Research team** creation, formation
- ➢ No meeting - **Thanksgiving**!

- ➢ Presentation 1: **US AI Labs #1** (During NeurIPs)
- ➢ Presentation 2: **US AI Labs #2**
- ➢ Presentation 3: **Chinese AI Labs**
- ➢ No meeting - **Christmas**!
- ➢ No meeting - **New Years Eve**!

- ➢ Presentation 4: Jan 6 - Conclusion of presentations