Finding the Right Abstractions Summit Summary

By J. Hirsh, with help from many

Introduction

This document is an attempt to abstract and summarize the May 2021 summit called *Finding the Right Abstractions* (*FRA*).

The summit grew out of fruitful exchanges between researchers at the Machine Intelligence Research Institute (MIRI), where they do foundational mathematical research to ensure smarter-than-human artificial intelligence has a positive impact, and Topos Institute, where they pioneer emerging mathematical sciences of connection and integration to steer humanity towards a better future.

There are many technical difficulties in Artificial Intelligence Safety (AIS), but one problem is that so many philosophical issues that are at its core—the meaning of agency, intelligence, flourishing, control—have not yet been abstracted well mathematically. While it is conceivable that domain-general AI might be achieved without solid mathematical foundations, it seems highly unlikely that this discovery will lead to human flourishing without that understanding in place.

Category theory has a history of providing big-picture insight into complex mathematical structures and their relationships, especially in situations where these structures are not amenable to descriptions by numbers or other simple invariants. Category theory is an approach to mathematics that emphasizes context over content, and it seems particularly well suited for abstracting the mathematics of some of the philosophical issues at the core of AIS.

This summit is intended to **cultivate that possibility**, by bringing together members of the AIS and ACT communities.

In the next few sections, I will report on what I learned and saw from attending the summit: its goals, organization, formal content, and informal content. I will reflect on how well these structures served the summit's goals, think through an emergent theme and some next possible steps.

Goals of the summit

1. Bring together the ACT community and AIS communities around the topics of X-risk and human flourishing.

2. Come away with a better sense of how the other field is thinking about these topics.

3. Identify possible collaborations and directions for future work.

Artificial Intelligence Safety (AIS)

Research scientists from Machine Intelligence Research Institute (MIRI), Center for Human-compatible Artificial Intelligence (CHAI), Future of Humanity Institute, and Pactum AI Inc. Their concerns include:

- How can we control the existential risk (X-risk) posed by AI?
- How can we ensure human beings maintain or improve current levels of control of the environment?
- How can we *mathematically prove* causal relationships exist using only statistical data? (Spoiler alert: sometimes you can, if you pay attention to how you've chopped up the world.)
- How can we ensure human flourishing?
- What is abstraction?
- What is a model?

Applied Category Theory (ACT)

Research scientists from Topos Institute, UC Riverside, Stanford University, University of Oxford, and Johns Hopkins University. Their concerns include:

- What is human? What is flourishing?

- What is intelligence?
- What is control?
- What is disagreement?
- What is a problem? What is a solution?
- What is a thing?
- What is a point of view?
- What makes something alive?

What happened

Summit Structure

Location

FRA took place via Gather, a video-calling space that combines video-calling with a 2D map, letting you move an avatar around and engage in video conversations with the other people whose avatars are close enough to yours. The virtual space included a stage from which the entire room could hear a participant's broadcasts, as well as private areas where one could be insulated from listening and hearing things going on outside the area.

Personnel

FRA included several greeters and an IT help desk for summit participants, as well as a moderator, videographer, and recorder.

Meetings

FRA met from 10am–1pm PDT on Tuesday and Wednesday for three weeks: May 4-5, 11-12, and 18-19. There were additional hours when the virtual space remained open for participants' use, including an occasion where a small group continued their discussion for more than five hours after the last lecture.

Lectures were either 20 or 40 minutes long, with a maximum of 3 lectures per day.

Resources

Before the meeting began, the organizers offered some videos to set the stage, including an introduction to Pure Category Theory and Applied Category Theory by David Spivak and a Fireside chat with Jaan Tallinn and David Dalrymple, which explores several central concerns of AI Safety, and identifies where in that space category theory might have a role to play.

The organizers prepared a summit Home Page google doc that participants could use as a central location for all summit-related pages, links, documents, etc.

The organizers prepared and shared google docs for each lecture, which they called Thought Catchers. These were used as repositories for questions and comments, and as collaborative workspaces for ideas that came up for the participants during lectures.

Lectures

Cultivating Strategies, David Spivak

While fully mature and uncontrolled AI would be quite dangerous, is it possible that we are already delegating great power to uncontrolled and immature intelligent processes? Distributed intelligence is a phrase that describes various kinds of intelligent processes, including the human mind, human cultures, computers, and corporations. A mathematical definition of 'cultivating strategies' would provide a foundation for reasoning about distributed intelligence.

Philosophy with a Deadline, Andrew Critch

How do we coordinate so that teams that produce AI breakthroughs are likely to have thought about the risks?

Three Realisms and the Idea of Sheaves, David Jaz Myers

A tour of modeling the world through *fixed realism, covariant realism, and local realism*. In fixed realism, there is just one model of the world; what is real is literally what the model says; this theory corresponds to the category of Sets. In covariant realism, there are many equivalent models of the world; what is real are the things that remain when passing from one equivalent model to the next; this theory corresponds to the category of Group Actions. In local realism, there are many inequivalent models of the world; what is real depends on **how disagreement is handled, which is part of the model**; this theory corresponds to the category of Sheaves, and disagreement possibilities are found in cohomology.

Truthful AI, Owen Cotton-Barratt

Truthful AI is a good target for human flourishing, and right abstractions are the bottleneck. What is truth, and does it have some underlying structure we can use to build truthful AI?

Abstraction = Information at a Distance - John S Wentsworth

Abstraction arises from information relevant "far away" in graphical probabilistic models. "Abstraction is a projection of a low-entropy Markov chain onto a low-entropy Markov chain, in such a way that the variables in the original model are independent when conditioned on the variables in the abstraction."

Finite Factored Sets, Scott Garrabrant

A finite factored set is categorically dual to a set with a partition; the factorization structure can be thought of as a way to decompose the set into concepts which determine the elements of the

set. The structure of a finite factored set is enough to recover much of Pearl's work on causal inference without taking variables or a casual graph as a given. In particular, given a finite factored set, one can produce a kind of causal graph, and *prove certain causal relationships exist by exhibiting certain kinds of statistical data*.

Information Geometry and Statistical Learning Theory, Alexander Oldenziel

Singular information geometry forms a powerful framework for machine learning.

Process Theory for Finding Right Abstractions, Toby St. Clere Smithe

How do things learn anything? A whirlwind tour of Toposes, Markov categories, Polynomial functors, and techniques for defining objects via their Toposes in the style of modern algebraic geometry. Introduced definitions for statistical games, providing insight into questions like: What does it mean to have a point of view? What is an action? What behaviors make things alive and how can those be modelled mathematically?

Resource Sharing Machines, Sophie Libkind

Dynamical systems abstract things that change; operads abstract their composition. Two particular kinds of compositional changing things: (1) *machines* take inputs, allow inputs to interact with their internal states, and provide outputs; two machines interact by wiring up inputs to outputs (2) *resource sharers* also take inputs, but they may share these with other resource sharers; resource sharers interact by both modifying the resources they share. Resource Sharing Machines are a categorical way to model both of these changing systems in the same mathematical setting. **Conjecture**: A "thing" is a changing system that interacts with other "things" in mostly "machine-y" ways.

Introduction to Categorical Logic, Evan Patterson

Classical logic is universal, but as such has very few models; some things it does model are overly complicated by this universality. Categorical logic gives a 'plug-and-play' toolkit for various forms of logic and reasoning.

Generalised Models as a Category, and Cartesian Frames, Stuart Armstrong

Models are strongly underdefined; can we build a mathematical framework that takes this into account? Generalized Models are an attempt to do this; what do they make easier to compute? What gets harder?

Automated Contract Negotiations, Kristjan Korjus

Large companies have to contract with thousands of suppliers/buyers/contractors, and even making all these deals can be incredibly costly and difficult to manage without waste. By working with the client to define their possible *contract space* and *value function*, Pactum AI,

Inc. is able to negotiate with the client's partners by making **maximally distinct counteroffers** with **equal value to the client**. In doing so, the client's are able to be flexible with their working partners while maximizing their own value.

Dialectica and Kolmogorov Problems, Valeria de Paiva

What is a problem, and what is a solution? How do these questions relate to intuitionistic logic? The category Dialectica and Chu objects model problems and answers in the style of Kolmogorov's "On the Interpretation of Intuitionistic Logic."

Symmetric Monoidal Categories: A Rosetta Stone, John Baez

Monoidal categories model concurrent interacting and **open** processes. "A cell phone is not a Turing Machine" because a Turing Machine starts with **one** input and is a deterministic process after that; a cell phone is **open**: constantly receiving inputs and producing outputs. Similarly, **intelligences**, **ecosystems**, and **organisms** are all open systems.

Nudges

In between lectures, the moderator would facilitate a Q&A session between the audience and speaker, and encourage interaction with the thought-catchers.

The moderator would also include short exercises designed to facilitate interaction and developing connections with new people. These included: reflecting on and sharing what you are bringing to the summit and what you hope to take away; sharing what conversations you are hungry for; planning small actions to make it more likely that we will continue to collaborate in the weeks to come.

The breakout-groups were often very light-touch: a brief session where people would get on stage to share what they hope to have a conversation about during the session, and then a short period of time where people mill about and find out who they will speak to about what.

Conversations

The organizers provided plenty of opportunity for the participants to interact in small groups within the virtual space. I was asked to leave room for private conversations, but I will include here a few beautiful lines I overheard.

> "If the agents try to maximize *the log of that which can be traded between them* then the boundaries between the agents don't matter so much; it's a kind of Gerrymander-proof goal." -conversation around Kelly Betting
>
> "Perhaps you should try to replace your notion of Truth with the structure of statements-and-evidence." -conversation around how to model Truth and Goodness for AI

Reflections

Goals

The summit seems to have been very successful at bringing together these two communities, creating opportunities for participants to learn about how their counterparts are thinking about the problems, and in fostering connections and possibilities for future work.

Summit Structure

Location

Gather was an excellent environment for the summit. There were some technical difficulties, and in the next iteration the organizers will probably make the virtual room bigger so that people can have more conversations that don't leak into each other, but otherwise this worked great.

Personnel

It was very useful to have greeters, and an IT help desk for the participants, and the organization went quite smoothly. The moderator was incredibly effective at directing participants' attention within the virtual space (to one another, to the Thought Catchers, to the speakers, and to the exercises). Many of the most innovative interaction ideas were guided by the moderator and these seemed quite fruitful. The videographer produced very high quality videos of the lectures and managed technical difficulties with the platform as they arose. [The following two sentences were written by D. Spivak:] The summit recorder felt like part of the group, but it was nice knowing that he had a specific role to play, and that his questions for clarification would lead to a better summary. It was extremely useful to have a philosophically-minded category theorist as our recorder.

Meetings

The lectures were shorter than at most math conferences, as were the day lengths, though the summit spanned a longer period of time than most conferences. This seemed to be an entirely positive change: shorter lectures and days make the summit accessible to a broader range of human attention spans (while still allowing those with more energy to continue on in optional sessions). The longer span of time for the summit also seemed to make space for the participants to get comfortable in the virtual environment, process what they were learning, and develop connections with fellow participants.

Resources

The introductions to pure and applied category theory were a great idea. The fireside chat exploring the central concerns of AI safety and its relation to category theory was an excellent contextualization of the goals of the summit.

Having a central Home Page was very useful and reduced organizational overhead for participants.

The Thought Catchers were also an excellent tool. Perhaps in the next iteration, there might be more Nudges from the moderator to interact with the Thought Catchers from past lectures in order to facilitate a return to those conversations and make the most of that resource.

Nudges

I have never been to a mathematics conference with so many thoughtfully-designed human-behavior exercises that facilitated participant interaction with each other and ideas and future collaborations. These were a breath of fresh air.

One feature that might be included in future iterations is some kind of explicit container for the breakout conversations. After all, conversations are distributed intelligences, and are subject to all kinds of dynamics, with some dynamics more suitable to serve the summit's goals. For example, some participants would politely mute their microphones when they were not speaking but other participants would leave their microphones unmuted (this distinction was strongly correlated with gender). As a result, I often witnessed a participant unmute their microphone to add something to a conversation, only to be talked over by someone who had left their microphone unmuted and was speaking quite frequently. Another conversational dynamic I witnessed that didn't seem very fruitful was a broader conversation becoming a dialogue between two participants, to the exclusion of the rest of the group.

Conversations: a theme emerges

As the summit recorder, I tried to move around and listen to as many conversations as possible. The topics of conversation were quite varied: mitigating X-risk, what is meaning, what is truth and good and how can we build AI to pursue truth and goodness, what is intelligence. In many of these conversations, I noticed some common features.

Often, someone working in AI safety would frame a question or propose a line of approach for dealing with one of these concepts. In response, someone (often a category theorist, but not always) would reply by problematizing the concept involved; that is, rather than working to solve the question posed, they would try to dissect the main concept at its core. I will include some abstractions of actual conversations I overheard:

> **AIS**: We want to build truthful and good AI, but how can we define the true and the good?
> **ACT**: Every model adds things to the world that aren't there, so what even is Truth?
> **AIS**: But some models are totally wrong! The earth isn't flat. Some models are more true than others.

ACT: Yes, definitely. Some models are strictly better than others. Some are better at some things and worse at others. No models are true.
AIS: How can I work on the problem of AI and truth while holding that paradox? It seems like in the examples I care about, I can make sense of what I mean by truth and goodness.

AIS: We want to control AI to mitigate X-risk.
ACT: Does a thermostat control the temperature of the room, or does the temperature of the room control the thermostat?

AIS: If an AI system achieves general intelligence, it will be quite dangerous to humans if its goals are not aligned with ours. Think about the paperclip maximizer.
ACT: Is a large corporation maximizing profits already a paperclip maximizing intelligence?

As is probably clear from my retelling, my philosophical sympathies are with the category theorists here. At the same time, it does seem a bit frustrating to be heckled in this way: to have one's premises problematized or presented with paradoxes while trying to tackle very important problems. And in most of the conversations I witnessed, I did not hear explicit attempts by the ACT crowd to shift from problematizing AIS concepts to providing tools with which to address the AIS concerns.

While I did not hear this shift explicitly in conversations, I did get the sense of what shape that answer might have by considering these conversations in the context of the summit lectures and what I know of how paradox has been handled by mathematicians in the past.

Russell's paradox, for example, is a problem with a theory of Sets if we assume that set construction is universally coherent. One traditional solution to this paradox is to insist that sets be defined **in context**. Rather than allowing set construction to be universally permitted, we introduce **new structure** in the theory: we incorporate permitted processes for making new sets from old ones. These new set-building structures protect the theory from Russell-like paradoxes by contextualizing the definitions of sets.

Banach-Tarski's paradox is a problem that arises in measure theory where we assume that measurability is universally coherent. One traditional solution to this paradox is to **contextualize** which sets can be measured by introducing **new structure** on a measurable space: the space must come equipped with an algebra of sets which are measurable.

Gödel's incompleteness theorems are a problem that arises in logical systems where we assume that every statement must be true or false. One modern solution to this paradox is to **contextualize** what it means to be true; we let go of the assumption that True and False are

universally coherent, that anything not-False is True, and instead direct attention to the structure of deductions.

In each of these cases, the mathematics in question was stymied by paradox arising from some **universalizing assumptions** about a concept (set containment, measurability, truth), and the solution in each case was to **get rid of the assumptions and replace them with structure that does the job**. If not all definitions make sets, what structures do we need to make sets? If not all sets are measurable, what structures do we need to work well with some collection of measurable sets? If not all not-False statements are provable, what structures do we need to understand which statements are provable?

Returning to the conversations described above, one summary of the ACT crowd's problematizations of the AIS concepts might be:

> *What universal assumptions about Truth, Intelligence, Control, can we replace with mathematical structures so that the important problems in AIS become easier to work with, and the paradoxes inherent in the universal versions fall away?*

This strategy showed up in several summit lectures. Scott Garrabrant gives up Pearl's universal variables and makes the way we have divided the world into variables part of his structure; in doing so he is able to develop a robust theory of causality and even learn some more subtle things about this causal structure than Pearl can. David Spivak presented the intuitions for what the structures of intelligence might look like, and how those structures can range in complexity from a single strategy to a collection of cultivated strategies acting as a distributed intelligence. John Baez described the mathematical structure of open systems and suggested that intelligence and agents should be modeled by open, rather than closed, systems. Toby St. Clere Smithe presented an ambitious mathematical portrait of what kinds of structures might describe an intelligent agent and its desires and actions.

To me, this conversational gap is both evidence that this summit was sorely needed, and also a pointer toward possibilities for future work. Perhaps it calls for another summit for *Testing the Right Abstractions*, where participants from both fields can work together to apply some of these abstractions and figure out whether any of them are good for the job.

Acknowledgments

Topos Institute thanks the following for making this summit possible:

> Funder: Centre for Effective Altruism, Long Term Future Fund
> Organizing Committee: David Spivak, Andrew Critch, and Scott Garrabrant
> Production team:

- Red Bridge Group, who provided administrative and a professional meeting organizing support
- Duncan Sabien, a talented facilitator with a deep understanding of human behavior and group process
- Braden Baumbach, a professional videographer,
- Joseph Hirsh, PhD, a category theorist who wrote updates for each day of the session and produced the final report