👋 Introduction

Welcome to the Exploring Explainables Reading Group!

MEETING CANCELED UNTIL FURTHER NOTICE — 2025.08.08

We use this document to keep track of readings, take notes during our sessions, and get more people excited about interactive scientific communication.

We have monthly reading groups on the ***second Friday of each month*** at ***10:30 am Eastern (US) Time***.

First time?

- Find the tab labeled with 🔜 for our **next session's meeting notes**, including the readings we will be discussing.
- To look at all the readings we've ever covered, check out the 📑 **Reading List** page.
- Head over to the 📇 **Contact** page to share your info and see who else has been involved.

What are *explainables*?

- *Explainables* are articles that leverage **dynamism** to convey complex concepts
- Other terms for this include
  - Interactive articles
  - Exploranations
  - Explorable explanations
  - etc.
- Here are some of our favorite sources for this kind of work:
  - Distill.pub
  - VISxAI Workshop
  - Explorable Explanations

Contact

Name	Background	Contact/ Socials	Interests
Shivam Raval		@sraval.bsky.social (bluesky), @jager.bomber (Discord), Linkedin. sraval@g.harvard.edu	See first reading group intro
Pranav Rajan		@pranavrajan.bksy.social (bluesky), @greenveggie320_(Discord). Pranav Rajan	Embeddings, High-Dimensional Visualization, Tooling, Interpretability,
Shreyans Jain	Independent AI Interpretability Researcher, 8 yrs of experience in Applied ML	@pyparrot.bsky.social (bluesky), @py_parrot (Twitter), @unnamed8802 (Discord), jshrey8@gmail.com	ML Interpretability, Reinforcement Learning, Good Visual Explainers of Complex Topics (main reason I liked distill)
Tyler Crosse	~8 years building web applications (mix of visualization and data engineering). Currently 1 year into an MSCS @ GA Tech	https://www.linkedin.com/in/tylercrosse/ Tyler Crosse - Georgia Institute of Technology | LinkedIn (add email)	Recently playing with SAE Lens and Neuronpedia. Thinking about the types of computation that happen in transformers.
Vishal Verma	~7 Years doing large scale ML and DS, Currently on break doing MS-ECE from CMU pittsburgh.	@v_shaal (X.com) vishalmverma27@gmail.com @vermavishal.bsky.social (bluesky) freeman0627 Discord https://www.linkedin.com/in/v-shaal/	Recently Started reading and working more on Mech Interp, Playing around tools to learn more. I am also exploring causal side of AI alignment.
YI-Wen Chu (William)	Currently 2 year into MSEE at Tatung University (Taiwan)	Email: william1006.chu@gmail.com Discord @williamsnail	Recently studying Mech Interpretability & AI Safety topics. New to this field, currently doing the ARENA curriculum, focus on SAE.
Jacob Haimes	Apart; Odyssean Institute, Kairos.fm (two podcasts)	• Bluesky: jacobhaimes.bsky.social • LinkedIn • muckrAIkers • Into AI Safety • Personal Website • Add email	Science communications, Improving institutional decision-making, LLM evals, bias mitigation, responsible & ethical AI

Chetan Talele	Undergrad; New to mech interp;Fellow at WhiteBox Research	- Twitter - Email - LinkedIn -	Working on benchmarking Theory of Mind;Doing the ARENA curriculum; Interested in agentic systems
Soumyadeep Bose	Pre-final Yr Undergrad | Fellow @WhiteBox Research | Researcher @AISC	email: soumyadeepboseee@gmail.com discord: @soumyadeepboseee linkedin: http://linkedin.com/in/soumyadeepbose/	
Rocco Di Tella	MSc in Data Science R.A neuroscience HMS R.A LLMs hallucinations	roccoditella1@gmail.com	
Seon Gunness	Independent ai researcher; datasci background, did MATS 2023 ; Been in ML for a while (2-3 years?) No uni background/unaffiliated	nseon103@gmail.com Disc: seonsmallworldz **(msg on disc; pmuch always active there)**	Mech interp, interpretability as a whole, explainability,both on a macro scale and math based scale (how small changes in tokens/training might affect stuff) Also I have a ton of links on discord/opportunities if you're still new/looking for stuff to do
Jason Rich Darmawan	MSc in Computer Science	Email: jasonrichdarmawan@gmail.com Discord: jasonrichdarmawan LinkedIn: https://linkedin.com/in/jasonrichdarmawan	Mech interp for Large Language Model

Aman Neelappa	Applied ML for 12+ years	Email: aman313@gmail.com	Interpreting and Steering LLMs

📑 Reading List

Year	Month	Reading(s)
2025	Feb	N/A
	March	Communicating with Interactive Articles and Research Debt
	April	A Gentle Introduction to Graph Neural Networks and Complexity Explained
	May	Feature Visualization
	~~June~~	Canceled
	July	A Visual Dive into Conditional Flow Matching

💾 2025.02.14 (February)

Before you go through the doc, I would like to share a quote I like

"You must be imaginative, strong-hearted. You must try things that may not work, and you must not let anyone define your limits because of where you come from. Your only limit is your soul. What I say is true - anyone can cook... but only the fearless can be great."
- Auguste Gusteau, Ratatouille (2007)

My intro

I'm Shivam Raval, PhD in Physics and AI from Havard (email sraval@g.harvard.edu)
Expected end date: fall-end 2025 (Dissertation plan approved! This is happening!)
Follow me on bluesky: @sraval.bsky.social
My LinkedIn: https://www.linkedin.com/in/shivam-raval-27820484/
Also on Discord: @jager.bomber
Wanna meet and chat? Book a meeting here

Thesis title (WIP): On creating visually interpretable explanations of high dimensional data

Some things I'm currently interested in and thinking about on the research side: quantum computing, interactive visualization, dimensionality reduction (pca / t-SNE / umap), manifold learning, topology and geometry, mechanistic interpretability approaches (patching / probing / SAEs), and visualization for scientific communication (eg. Distill and explorables), full-stack pipelines for foundation models, AI Safety and alignment, ai4physics, systems4ml, systems biology, and neuroscience, and psychology.

Some slide decks from previous presentations:
1. How to do good research?
2. Thinking in analogies
3. Specific topics: SAEs, Hypertrix (Got the best paper award)

Distill: Intro meet

Feb 14, 11.30am ET - 1pm ET. Meeting link here

Sharing resources

1. https://distill.pub/

2. https://visxai.io/
3. https://pair.withgoogle.com/explorables/
4. https://www.fastht.ml/
5. https://www.answer.ai/posts/2025-01-15-monsterui.html
6. https://mlu-explain.github.io/
7. https://poloclub.github.io/
8. https://transformer-circuits.pub/
9. https://www.neuronpedia.org/
10. https://www.techwithtim.net
11. Here's an example of an article I provided feedback on, I think it's really effective (best thing I've found) on explaining RLH(AI)F, and is targeted at an interested layperson: https://kairos.fm/simple-technical-rlhaif/
12. https://www.apartresearch.com/post/hunting-for-ai-hackers-in-the-wild-llm-agent-honeypot
13. https://playground.tensorflow.org/ - Can visualize effect of activation function on regression,layers,neurons etc. Not specific to LLMs but a part of explainable AI
14. Transformer Explainer: LLM Transformer Model Visually Explained
15. Andy Matushak's mnemonic book on quantum computing

Reading List

1. https://distill.pub/2020/communicating-with-interactive-articles/
2. https://distill.pub/2017/research-debt/

Challenges

- Shivam: good tooling, svelte vs react vs js vs R etc
- Pranav: Distill is very hard to use if you're unfamiliar with vanilla web development and d3
- Jacob: Resources on tooling (starter guide); getting other people to do it; time; the path (where to start, what to understand before other things)
- Pranav: No clear solution python interactive visualization solution for machine learning (pysvelte by anthropic) - +100 on this (shreyans)
- Design skills. Making visualizations that are easy to understand and that look nice is difficult. Dedicated designers often lack technical understanding, and technical people often don't have design expertise.
- Shreyans: Linear algebra
- Jacob: a path to mech interp research
    - This was helpful for me:
    - https://www.neelnanda.io/mechanistic-interpretability/getting-started
    - ARENA? Various fellowships?

Potential solutions

1. Some sort of user study/test run to understand what kind of mode of communication works for which demographics and folks of different backgrounds (Rocco: andy matushak, ask questions throughout the article)
2. Rocco: I think the pen and paper version of a good post is where a lot of the work has to go into. Choosing a single good interactive element can be totally enough.
3. Soumyadeep: Cool math book I talked about: https://mml-book.github.io/book/mml-book.pdf (it's intense though)
4. Different paths for folks with different backgrounds
    a. Physics and math
    b. CS
    c. Non physics and math background but knows how to code
    d. Non-physics and math background but does not know to code
5. Jabob: footnotes to articles that go deeper into the concept
    a. Examples:
        i. Here's an example of an article I provided feedback on, I think it's really effective (best thing I've found) on explaining RLH(AI)F, and is targeted at an interested layperson → https://kairos.fm/simple-technical-rlhaif/
        ii. About detecting AI hackers, requires understanding of prompt injection → https://www.apartresearch.com/post/hunting-for-ai-hackers-in-the-wild-llm-agent-honeypot
6. Chetan : Dumbed-down glossary of words being hyperlinked each time certain concepts are mentioned.Can avoid putting newbies into rabbitholes
7. Rocco: Andy Matushak's mnemonic book on quantum computing
8. Overview of the landscape of data visualization tools.
    a. How does d3 relate to javascript / React / svelte / web sites ?
    b. Where do python tools like matplotlib, plotly, streamlit, etc. fit in?
    c. How do you render latex, syntax highlighted code, etc.?
    d. Pros/cons of static images vs. interactive visuals vs. videos
    e. Question to answer → what's the minimum I need?

Brainstorming potential Distill-like ideas

1. Jacob: Starter guide on Distill, visual abstracts/Teaser images/Hero image
2. Shivam: SAEs and Quantum Physics
3. Pranav: SAEs, Benchmarking

4. Jacob: visual abstracts; how to write a blog; SVM(?), linear algebra, DMDU/RDM/DAPP
5. Shreyans: Understanding basics of mech interp - basis, privileged basis, superposition
6. Vishal : More Manim (3b1b) style animation for explaining maths behind Interp
7. Chetan : Develop visualizations that trace the flow of information from input to output
8. Seon: Transformer-lens
9. Yi-Wen: I think we need a explaining video similar to this Map of Computer Science to make people more clear on the path about what to do on Mech Interp research.

10. Shreyans: Reinforcement Learning Algorithms? (Is it something which can be considered as well?)(all the images generally used are toy examples, nothing to explain the complex algos)

Agenda

1. Introduction (~40mins):
    a. Opening remarks (5mins)
    b. Everyone puts down their contact details, interests, resources, and challenges to continuing Distill style work (5min)
    c. Introduction: 30mins (if 60 participants, then everyone gets 30 seconds)
2. Socialization (~35 + 5 minute break):
    a. Randomized 1 (10min): 3-5 people in a breakout room
    b. Based on interests (15mins): 3-5 people in a breakout room
    c. Randomized 2 (10min): 3-5 people in a breakout room
3. Closing Remarks:
    a. Putting down potential solutions to challenges (8 mins)
    b. End note (2mins)

💾 2025.03.14 (March)

APOLOGIES!!!

The google calendar event was deleted by accident ~1 hour before the event, which caused some confusion. I will make sure this doesn't happen next time.

If there is ever any confusion, and/or you want to be added to my Google Calendar event for this meetup, please reach out to `Jacob Haimes` and include a little bit of context — Thanks in advance!

Readings

Communicating with Interactive Articles (proposed by Jacob) and Research Debt (proposed by Rocco)

Agenda

- Brief intros
- Summary of articles - Jacob & Rocco
- Discussion of articles
- Choose next reading
- Projects

Discussion

Please add points you'd like to discuss during the meetup, including your favorite parts, how to replicate certain visualizations, potential improvements, skepticism, questions, points of confusion, or anything else relating to the reading.

- Maybe it is quite hard for someone good at academics to get into the interactive visualization technique?
- How do we make the incentives better for this kind of work?
    - Maybe in human centered design would be amenable to workshops

- Delta between interactive vs. videos like 3Blue1Brown
- What about Bacteria Evolution Simulation
    - Maybe uptake is related to awareness?
    - Diffusion of Innovation Theory | Canadian Journal of Nursing Informatics
- What's the difference between interactivity and just visualizations
    - https://www.complexityexplorer.org/explore/virtual-laboratory
    - Just visualizations hit (i) and (v), but not the others
        i. Connecting People and Data
        ii. Making Systems Playful
        iii. Prompting Self-Reflection
        iv. Personalizing Reading
        v. Reducing Cognitive Load
    - Out of the three remaining, which is the most important?
        i. Jacob: maybe prompting self-reflection(?) if we are aiming for understanding
        ii. Chetan: Game-ifying also helps people want to learn how to do best
    - Sruthi: Monitor: An AI-Driven Observability Interface | Transluce AI & Transluce Monitor - Tools like Monitor and FoldIt can drive research

Choosing our next reading

If you have an article you think would be worth reading in this group, include a link to it in the table below, and be ready to give a <60 second pitch on why we should read it!

Proposed Reading	Votes (conducted during meeting)
A Gentle Introduction to Graph Neural Networks	3
Complexity Explained	2

Explorable Explanations	0

For future reference, some good interpretability articles are:

- Feature Visualization
    - For more background on DeepDream, there's this video
- Zoom In: Intro to Circuits
- Toy Models of Superposition (long)

Projects

Please add anything you'd like to bring up after the discussion of the reading. This can include current projects, asking for specific advice, discussing a particular pain-point, providing resources, sharing new findings and opportunities, etc.

-

💾 2025.04.11 (April)

Readings

A Gentle Introduction to Graph Neural Networks (Sruthi) and Complexity Explained (Jacob)

Agenda

- Brief intros
- Summary of articles - Sruthi & Jacob
- Discussion of articles
- Choose next reading
- Projects

Discussion

Please add points you'd like to discuss during the meetup, including your favorite parts, how to replicate certain visualizations, potential improvements, skepticism, questions, points of confusion, or anything else relating to the reading.

- Two points only one demo — dynamics
- Phase transitions could have been its own section?
    - Show water molecules and what it looks like
    - More is different → linked to broken symmetry, when something happens that breaks the possibility to be symmetrical
- Organization of it: some of the concepts could have been grouped a little bit differently, a little bit more divided
    - Maybe focus on one aspect and then go back and show how that connects to the others
    - Foundational Papers in Complexity Science
    - The Complex World: An Introduction to the Foundations of Complexity Science — SFI Press

- Ants!
    - Ant problem: ant trap uses borax and hijacks the patterns against them

Choosing our next reading

If you have an article you think would be worth reading in this group, include a link to it in the table below, and be ready to give a <60 second pitch on why we should read it!

Proposed Reading	Votes (conducted during meeting)
https://distill.pub/2017/feature-visualization/	

Projects

Please add anything you'd like to bring up after the discussion of the reading. This can include current projects, asking for specific advice, discussing a particular pain-point, providing resources, sharing new findings and opportunities, etc.

-

Message from Shivam

Hi all, sorry I could not attend this meeting, but I'm thinking of working on some Distill-style interactive visual articles (topics TBD, but can include interactive blogs on complex mathy topics like subspaces and manifolds, how reasoning models work, interp approaches like probing/SAEs/Crosscoders/CLTs, etc.). If this interests you, please feel free to reach out to me (contact details on the contact page)

💾 2025.05.09 (May)

Readings

Feature Visualization (Morgan)

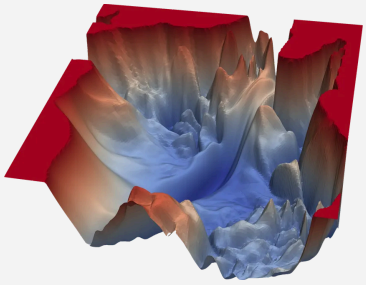
Agenda

- Brief intros
- Summary of articles - Morgan
- Discussion of articles
- Choose next reading
- Projects

Discussion

Please add points you'd like to discuss during the meetup, including your favorite parts, how to replicate certain visualizations, potential improvements, skepticism, questions, points of confusion, or anything else relating to the reading.

- Consistent and similar to concepts from neuroscience
- Basic idea: instead of training your NN as you usually would, you instead optimize to the image, and you maximize the activation of a particular neuron
    - This freezes the NN, and the pixels are the features
    - You modify the pixels to maximize the activations of a specific feature
    - People have started doing this for LLMs now as well
- Multiple approaches to baseline test whether the output images actually represent something within the data – they just sort of visually verified w/ "dataset exemplars"
- Initial conditions of the optimization impact the output image that you reach
    - https://arxiv.org/pdf/2402.10039

    - Loss landscape: 
- Class probability
    - You end up w/ logits (how much evidence for each number) → when the softmax is applied, everything has to add up to 1, the difference between *class logits* and *class probability* is that class logit optimizes a specific classification, but the class probability also effectively minimizes the other classes
- Have tried doing this for language →there is some success now
    - What does BERT dream of?
    - Fluent dreaming for language models
- A criticism is that you're basically doing a Rorschach test, so it's possible to be implying something that isn't there
- Similar patterns occur in human neurons → maybe where some of the information comes from
    - For visual processing in animals, there are a few feedforward layers in neurons; neurons that are "closer" to the eye (input) are analogous to edge detectors (gabor filters), later on you have neurons starting to codify higher level concepts
    - Note: CNNs also respond very strongly to gabor filters
    - An Overview of Early Vision in InceptionV1 → Gabor Filters
- Sometimes it feels like single examples are over-explained in the article
    - Without a rigorous baseline it doesn't really make sense to conjecture about less obvious patterns within the images
-

Choosing our next reading

If you have an article you think would be worth reading in this group, include a link to it in the table below, and be ready to give a <60 second pitch on why we should read it!

Proposed Reading	Votes (conducted during meeting)
AI 2027	0
A Visual Dive into Conditional Flow Matching | ICLR Blogposts 2025	4

Projects

Please add anything you'd like to bring up after the discussion of the reading. This can include current projects, asking for specific advice, discussing a particular pain-point, providing resources, sharing new findings and opportunities, etc.

-

💾 2025.07.11 (July)

Readings

A Visual Dive into Conditional Flow Matching (Shivam)

Agenda

- Brief intros
- Summary of articles - Shivam
- Discussion of articles
- Choose next reading
- Projects

Discussion

Please add points you'd like to discuss during the meetup, including your favorite parts, how to replicate certain visualizations, potential improvements, skepticism, questions, points of confusion, or anything else relating to the reading.

-

Choosing our next reading

If you have an article you think would be worth reading in this group, include a link to it in the table below, and be ready to give a <60 second pitch on why we should read it!

Proposed Reading	Votes (conducted during meeting)

Projects

Please add anything you'd like to bring up after the discussion of the reading. This can include current projects, asking for specific advice, discussing a particular pain-point, providing resources, sharing new findings and opportunities, etc.

-

💿 Template

Readings

Agenda

- Brief intros
- Summary of articles -
- Discussion of articles
- Choose next reading
- Projects

Discussion

Please add points you'd like to discuss during the meetup, including your favorite parts, how to replicate certain visualizations, potential improvements, skepticism, questions, points of confusion, or anything else relating to the reading.

-

Choosing our next reading

If you have an article you think would be worth reading in this group, include a link to it in the table below, and be ready to give a <60 second pitch on why we should read it!

Proposed Reading	Votes (conducted during meeting)

Projects

Please add anything you'd like to bring up after the discussion of the reading. This can include current projects, asking for specific advice, discussing a particular pain-point, providing resources, sharing new findings and opportunities, etc.

-