Mechanistic?

Here's the image representing "mechanistic interpretability" with a network of interlocking gears, circuits, and nodes lighting up to convey understanding within a mechanical and neural structure. Let me know if you'd like any adjustments or additional details!

DALL-E 3 rendition, 11/14/24

Naomi Saphra* and Sarah Wiegreffe*

arxiv.org/abs/2410.09087

Mechanistic Interpretability is extremely hot

Mechanistic Interpretability is extremely hot

Sasha Rush ✓
@srush_nlp

I recently asked pre-PhD researchers what area they were most excited about, and overwhelmingly the answer was "mechanistic interpretability". Not sure how that happened, but I am interested how it came about.

Last edited 10:11 AM · Jan 23, 2024 · **284.5K** Views

Mechanistic Interpretability is extremely hot– but what is it?

Sasha Rush
@srush_nlp

I recently asked pre-PhD researchers what area they were about, and overwhelmingly the answer was "mechanistic interpretability". Not sure how that happened, but I am inte came about.

Last edited 10:11 AM · Jan 23, 2024 · **284.5K** Views

Jacob Andreas
@jacobandreas

I still don't totally understand the difference between "mechanistic" and "non-mechanistic" interpretability but it seems to be mainly a distinction of the authors' social network?

5:11 PM · Jan 23, 2024 · **10.9K** Views

Mechanistic Interpretability is extremely hot– but what is it?

Sasha Rush
@srush_nlp

I recently asked pre-PhD researchers what area they were about, and overwhelmingly the answer was "mechanistic interpretability". Not sure how that happened, but I am inte came about.

Last edited 10:11 AM · Jan 23, 2024 · **284.5K** Views

Jacob Andreas
@jacobandreas

I still don't totally understand the difference between "mechanistic" and "non-mechanistic" interpretability but it seems to be mainly a distinction of the authors' social network?

5:11 PM · Jan 23, 2024 · **10.9K** Views

Nathan Benaich
@nathanbenaich

is mechanic interpretability a sexier way of saying interpretability?

5:52 AM · Jul 28, 2024 · **11.9K** Views

Mechanistic Interpretability is extremely hot– but what is it?

Sasha Rush
@srush_nlp

I recently asked pre-PhD researchers what area they were about, and overwhelmingly the answer was "mechanistic interpretability". Not sure how that happened, but I am inte came about.

Last edited 10:11 AM · Jan 23, 2024 · **284.5K** Views

Nathan Benaich
@nathanbenaich

is mechanic interpretability a sexier way of saying int

5:52 AM · Jul 28, 2024 · **11.9K** Views

Jacob Andreas
@jacobandreas

I still don't totally understand the difference between "mechanistic" and "non-mechanistic" interpretability but it seems to be mainly a distinction of the authors' social network?

5:11 PM · Jan 23, 2024 · **10.9K** Views

Andrew Gordon Wilson
@andrewgwils

Did they seem to know much about it and the foundations? I've also noticed a major increase in interest in this area, and alignment, but I suspect unfortunately for many it's just trendy buzzwords.

8:11 PM · Jan 23, 2024 · **1,866** Views

Timeline

Interpretability picks up in deep learning, NLP, vision

2015-2017

present

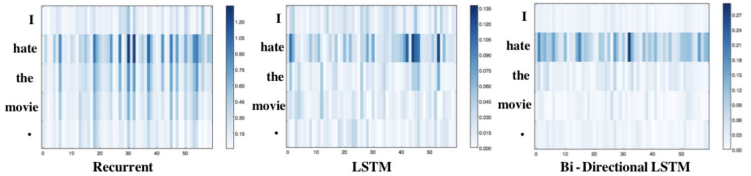


Figure 5: Saliency heatmap for for "I hate the movie ." Each row corresponds to saliency scores for the correspondent word representation with each grid representing each dimension.

Li et al, 2016

Timeline

First BlackBoxNLP workshop

Interpretability picks up in deep learning, NLP, vision

2015-2017

Nov 2018

present

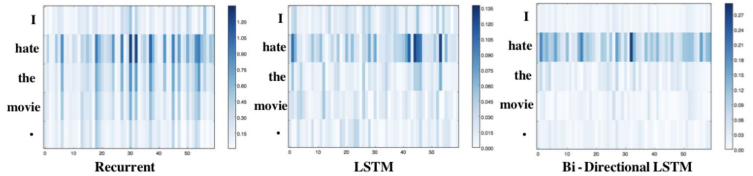


Figure 5: Saliency heatmap for for "I hate the movie ." Each row corresponds to saliency scores for the correspondent word representation with each grid representing each dimension.

Li et al, 2016

Timeline

First BlackBoxNLP workshop

Interpretability picks up in deep learning, NLP, vision

2015-2017

Nov 2018

present

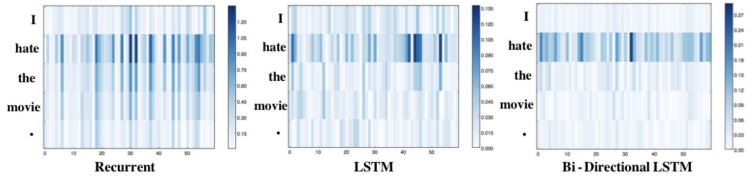


Figure 5: Saliency heatmap for for "I hate the movie ." Each row corresponds to saliency scores for the correspondent word representation with each grid representing each dimension.

Li et al, 2016

Timeline

First BlackBoxNLP workshop

First Interp + Analysis track at *CL conference

Interpretability picks up in deep learning, NLP, vision

2015-2017

Nov 2018

July 2020

present

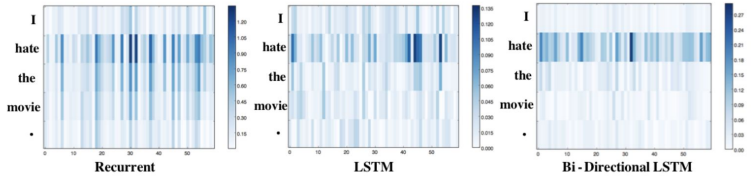


Figure 5: Saliency heatmap for for "I hate the movie ." Each row corresponds to saliency scores for the correspondent word representation with each grid representing each dimension.

Li et al, 2016

Overview: 4 Definitions

1. **Narrow technical:** understanding neural networks through causal mechanisms implemented by their internal components
2. **Broad technical:** any research describing the internals of a model
3. **Narrow cultural:** any research originating from the MI community
4. **Broad cultural:** any research in the field of AI—especially LM—interpretability

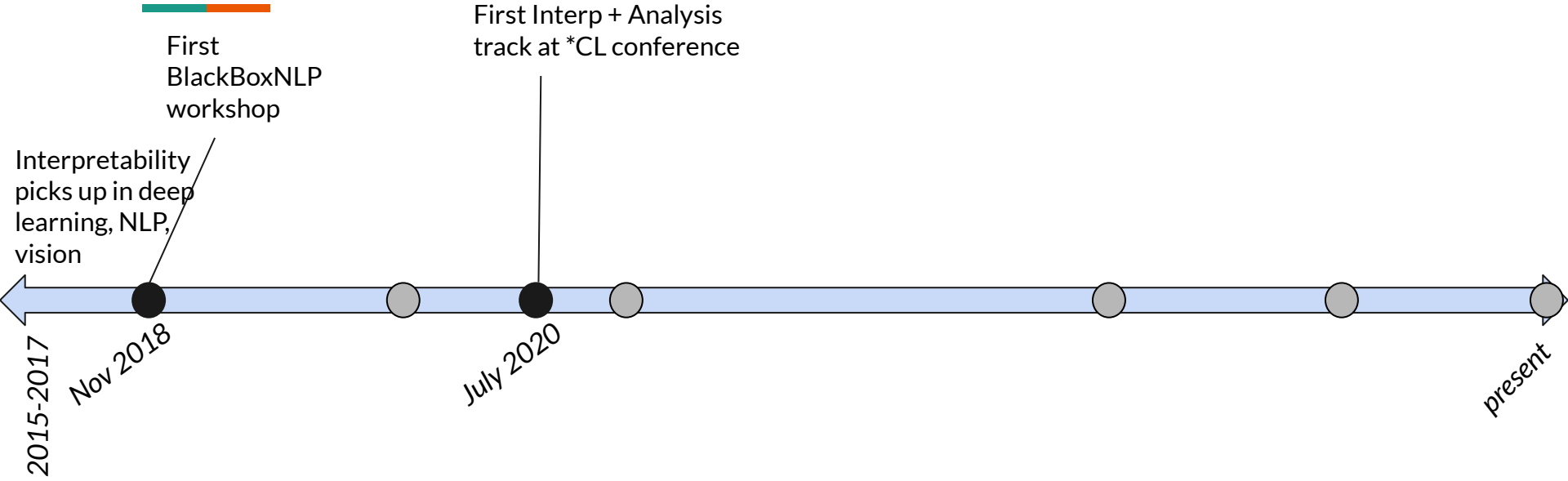
Overview: 4 Definitions

1. **Narrow technical:** understanding neural networks through causal mechanisms implemented by their internal components
2. **Broad technical:** any research describing the internals of a model
3. **Narrow cultural:** any research originating from the MI community
4. **Broad cultural:** any research in the field of AI—especially LM—interpretability

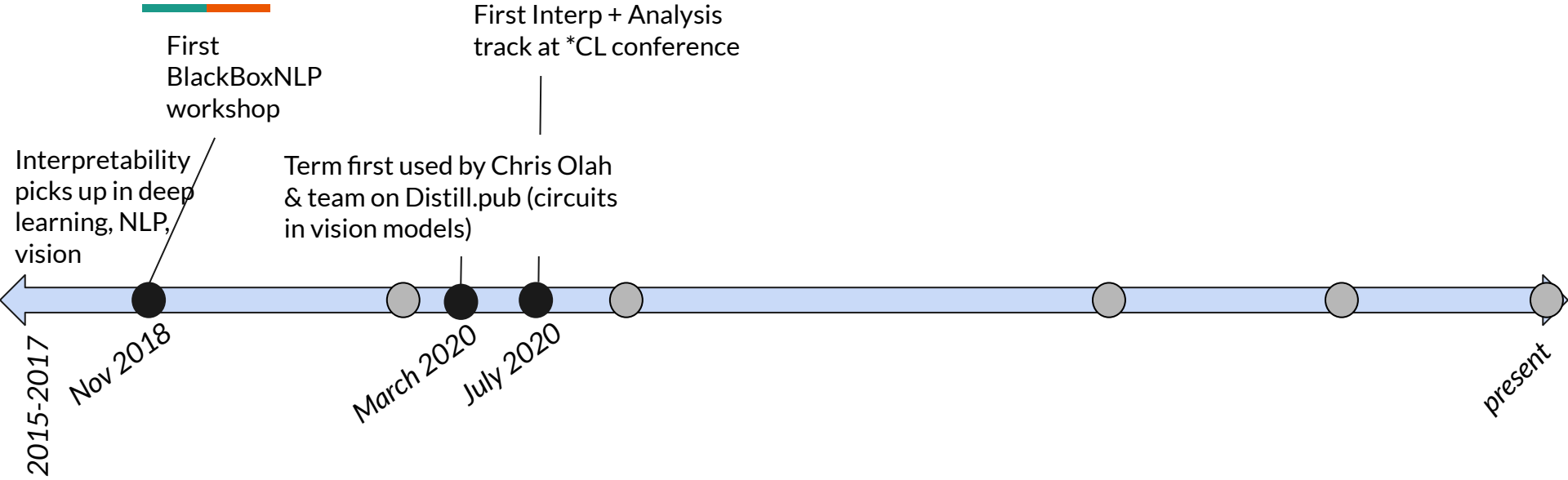
Overview: 4 Definitions

1. **Narrow technical:** understanding neural networks through causal mechanisms implemented by their internal components
2. **Broad technical:** any research describing the internals of a model
3. **Narrow cultural:** any research originating from the MI community
4. **Broad cultural:** any research in the field of AI—especially LM—interpretability

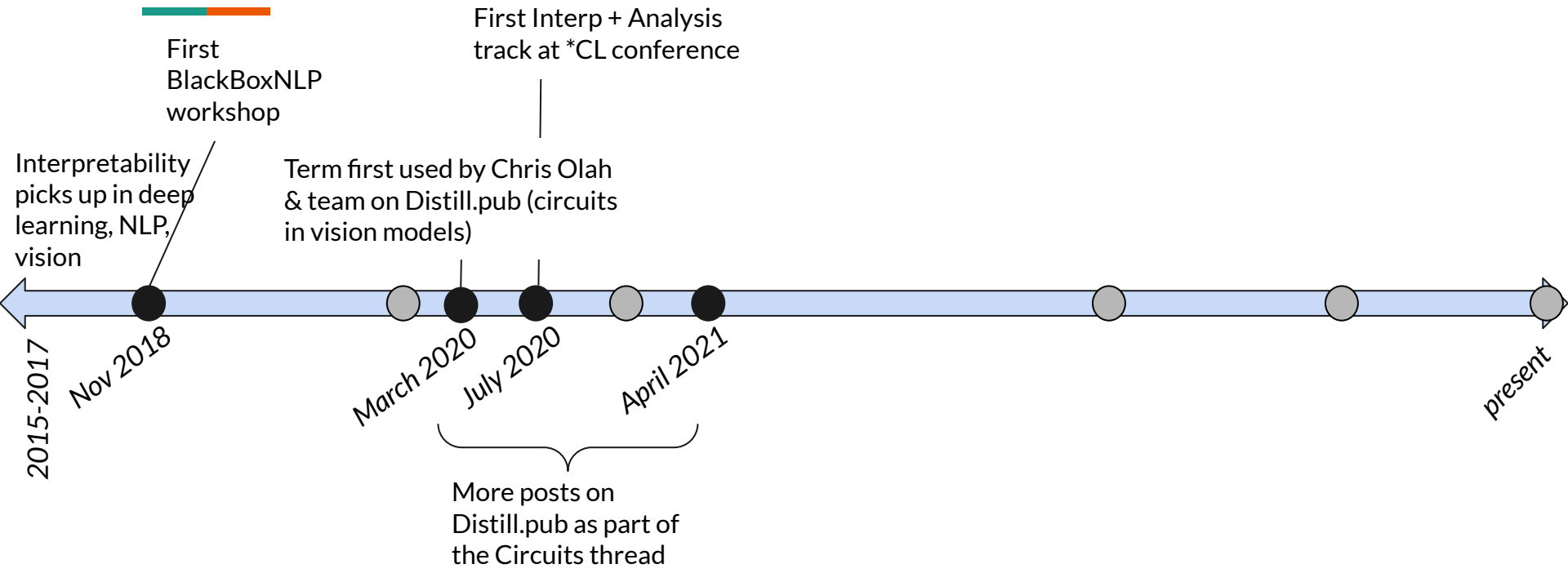
Timeline


First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Interpretability picks up in deep learning, NLP, vision
2015-2017
Nov 2018
July 2020
present

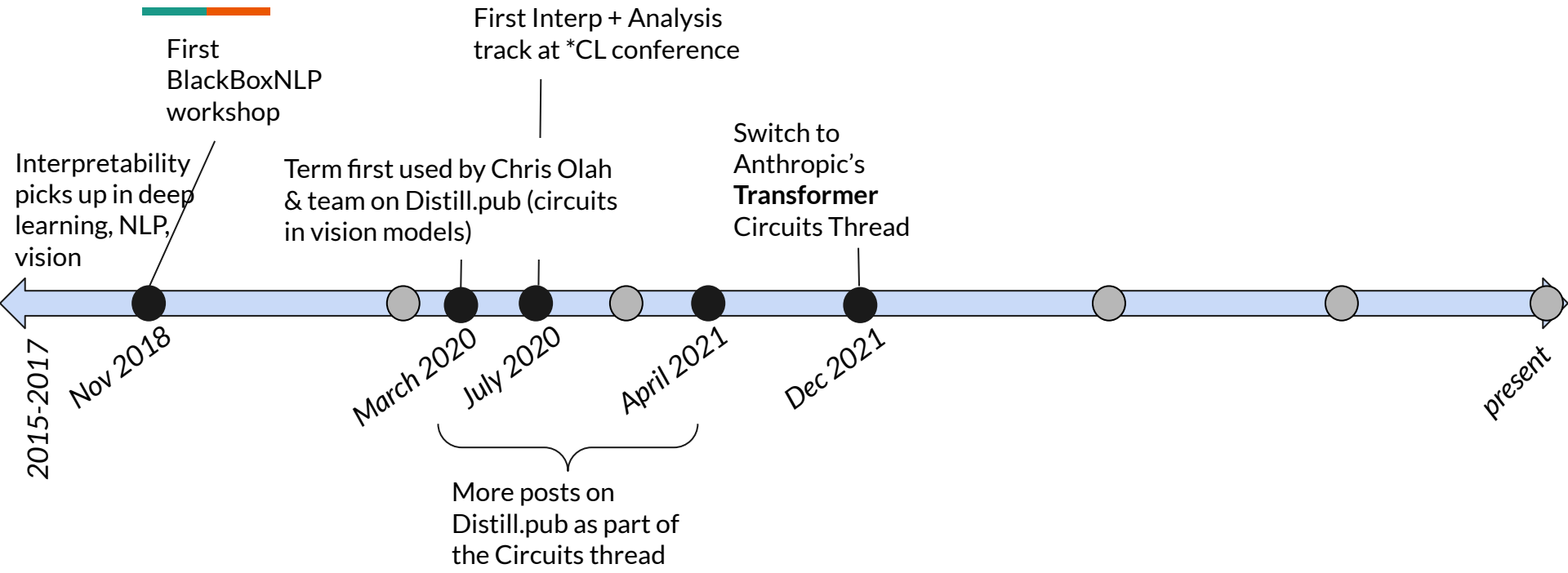
Timeline


First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
2015-2017
Nov 2018
March 2020
July 2020
present

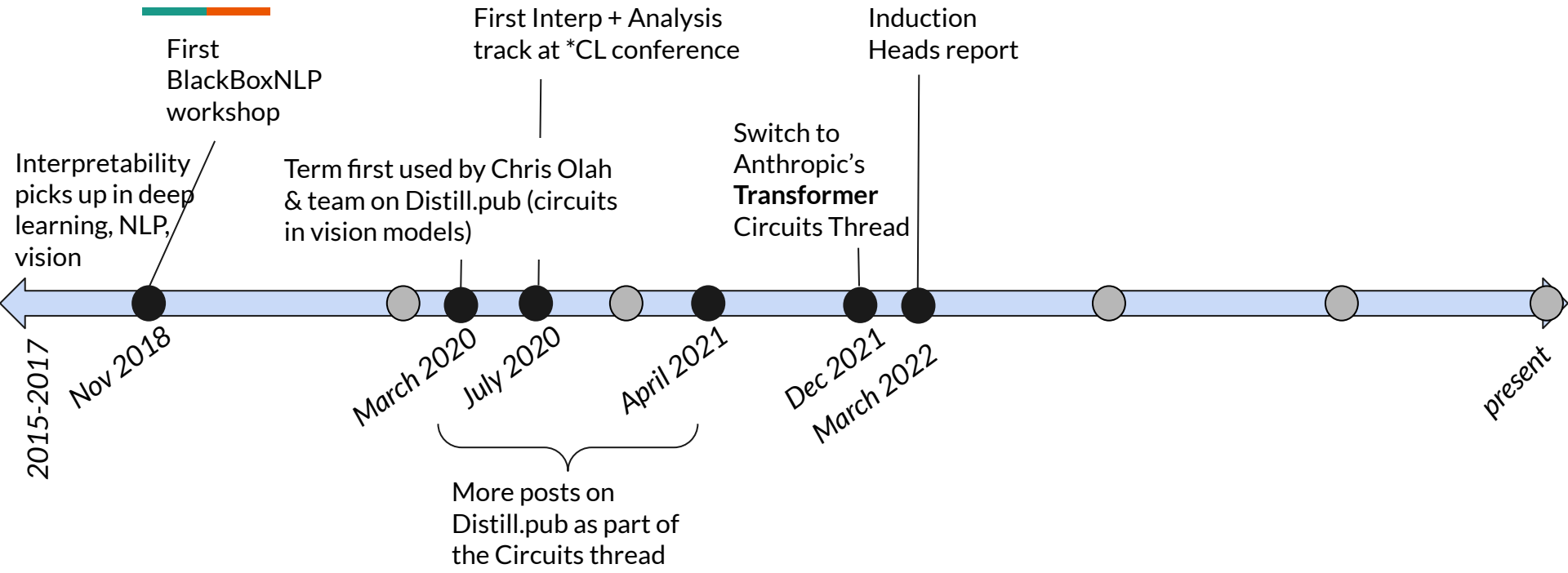
Timeline


First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
2015-2017
Nov 2018
March 2020
July 2020
April 2021
present
More posts on Distill.pub as part of the Circuits thread

Timeline

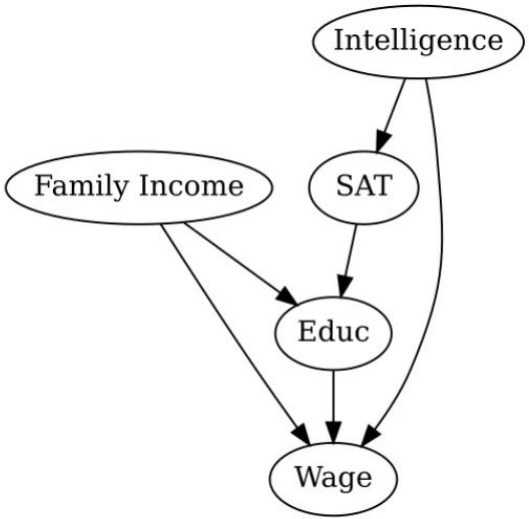

First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
Switch to Anthropic's **Transformer** Circuits Thread
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
present
More posts on Distill.pub as part of the Circuits thread

Timeline


First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Induction Heads report
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
Switch to Anthropic's **Transformer** Circuits Thread
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
present
More posts on Distill.pub as part of the Circuits thread

Causality is key!

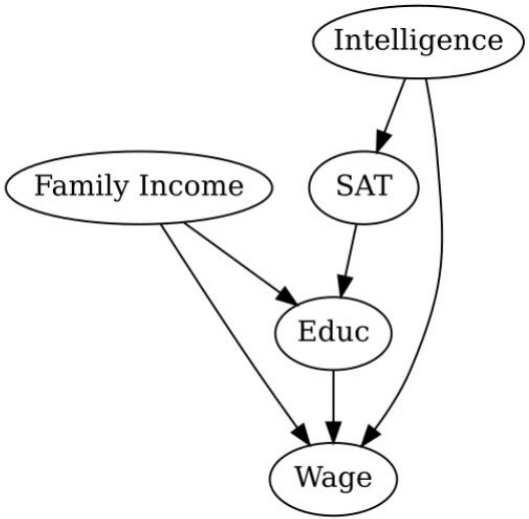
- Causal models are made up of **causal mechanisms**
- Causal mechanism = a function that transforms some subset of model variables (causes) into another subset (outcomes or effects)
    - See the analogy?



Halpern & Pearl, 2005

Causality is key!

- Causal models are made up of **causal mechanisms**
- Causal mechanism = a function that transforms some subset of model variables (causes) into another subset (outcomes or effects)
    - See the analogy?

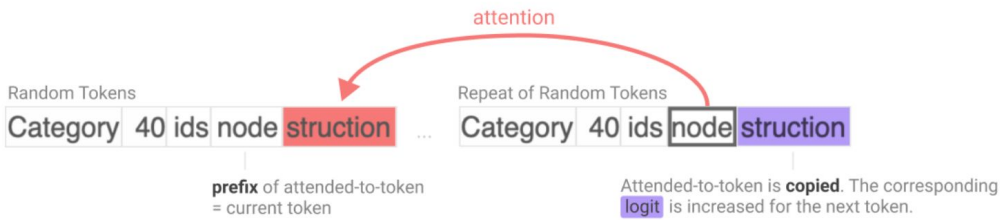


MI = research that discovers causal mechanisms explaining **all or some part** of the change from neural network input to output at the level of intermediate model representations

Halpern & Pearl, 2005

Causality is key!

- Example: induction heads
  - Question: **how does LM predict “B” given “ABABA”?**
  - Answer:
    - Attention heads search for a previous occurrence of “A”
    - Other heads then attend to the token that follows it



Olsson et al, 2021

Causality is key!

- Recent work proposes for an **even narrower definition**: characterizing a **complete, end-to-end** pathway from model inputs to outputs
- This definition isn't yet widely agreed upon
    - In part because it excludes work like Induction Heads

Geiger et al., 2021, 2024b; Mueller et al., 2024

Causality is key!

- Recent work proposes for an **even narrower definition**: characterizing a **complete, end-to-end** pathway from model inputs to outputs
- This definition isn't yet widely agreed upon
    - In part because it excludes work like Induction Heads

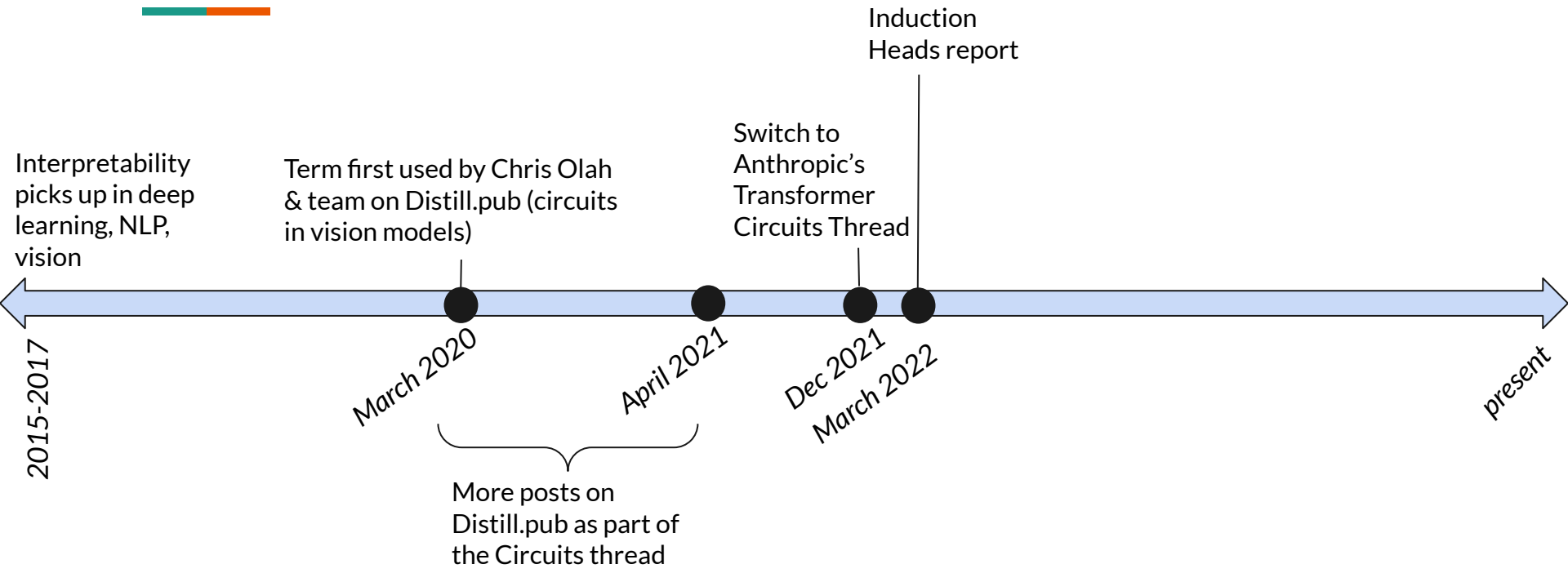
MI = research that discovers causal mechanisms explaining **all ~~or some part~~** of the change from neural network input to output at the level of intermediate model representations

Geiger et al., 2021, 2024b; Mueller et al., 2024

Overview: 4 Definitions

1. **Narrow technical:** understanding neural networks through causal mechanisms implemented by their internal components
2. **Broad technical:** any research describing the internals of a model
3. **Narrow cultural:** any research originating from the MI community
4. **Broad cultural:** any research in the field of AI—especially LM—interpretability

Timeline


Induction Heads report
Switch to Anthropic's Transformer Circuits Thread
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
2015-2017
March 2020
April 2021
Dec 2021
March 2022
present
More posts on Distill.pub as part of the Circuits thread

Coinage of Mechanistic Interpretability

"... reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by **examining the weights and activations** of neural networks to identify circuits ... that implement particular behaviors."

Olah et al. 2020, Elhage et al. 2021, MI workshop

Coinage of Mechanistic Interpretability

"... reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by **examining the weights and activations** of neural networks to identify circuits ... that implement particular behaviors."

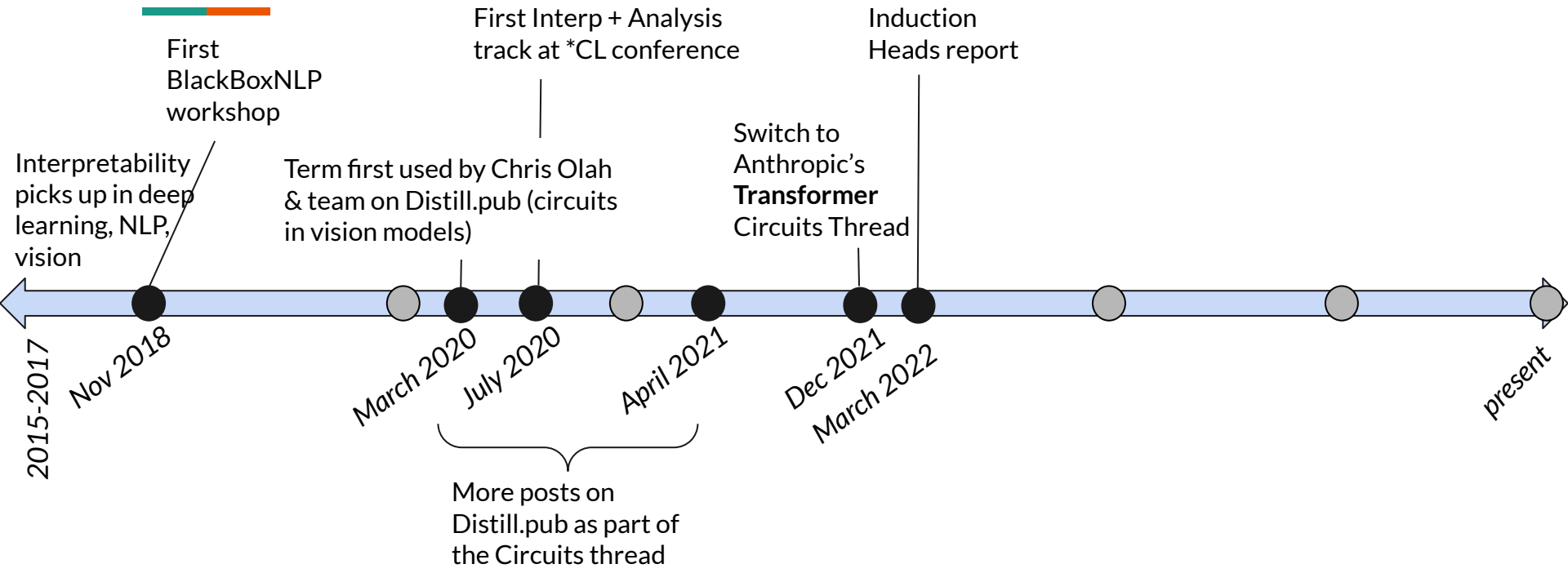
MI = *any* inspection of intermediate model representations or weights

Olah et al. 2020, Elhage et al. 2021, MI workshop

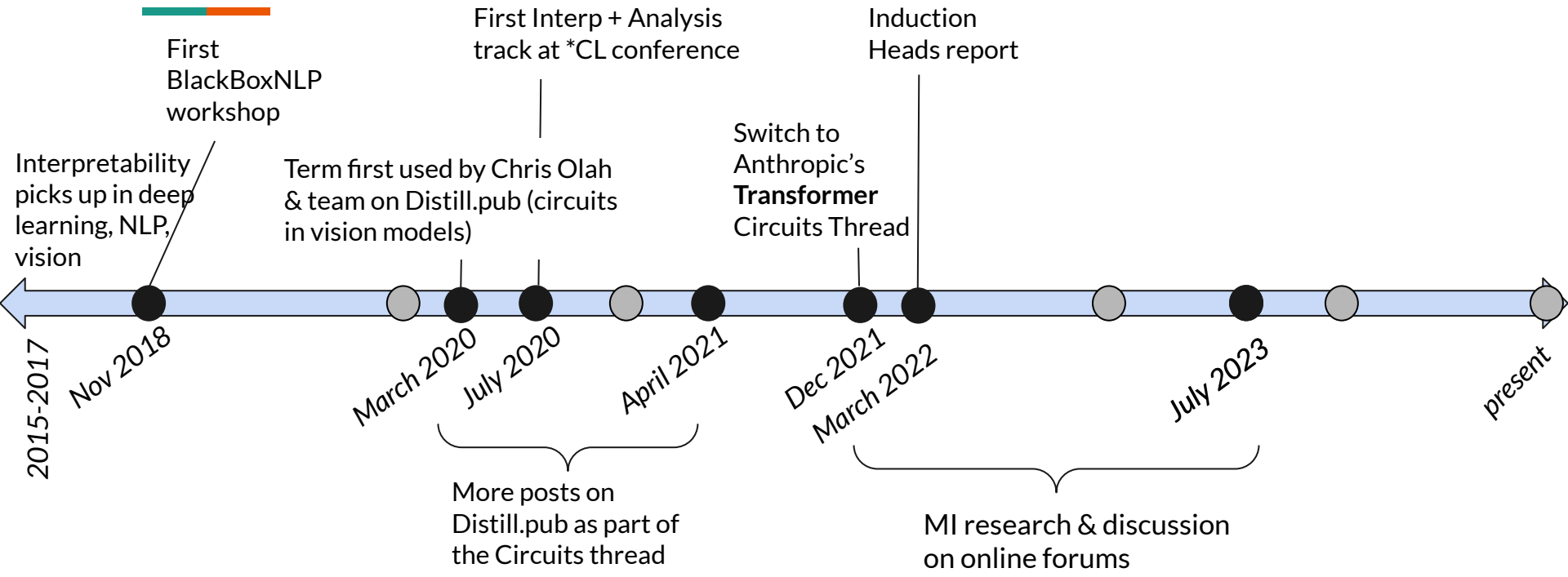
Overview: 4 Definitions

1. **Narrow technical:** understanding neural networks through causal mechanisms implemented by their internal components
2. **Broad technical:** any research describing the internals of a model
3. **Narrow cultural:** any research originating from the MI community
4. **Broad cultural:** any research in the field of AI—especially LM—interpretability

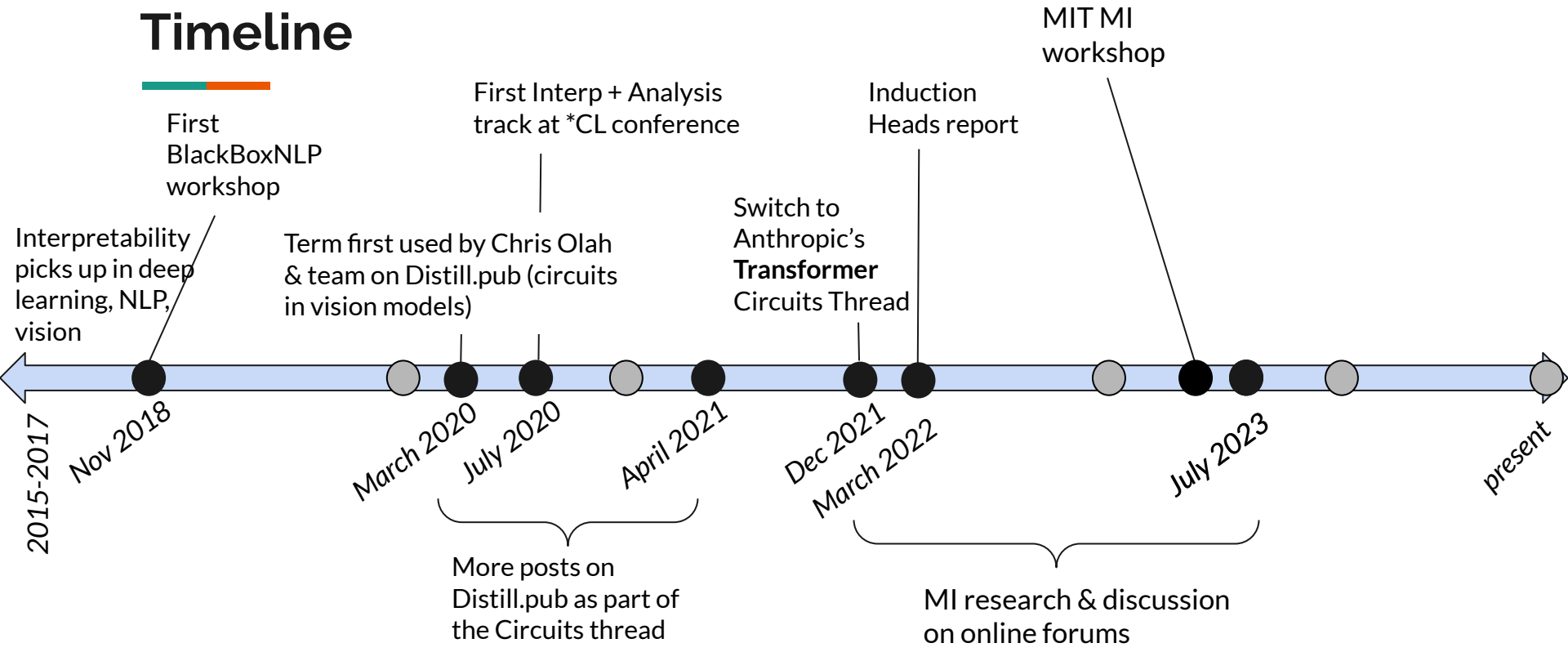
Timeline


First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Induction Heads report
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
Switch to Anthropic's **Transformer** Circuits Thread
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
present
More posts on Distill.pub as part of the Circuits thread

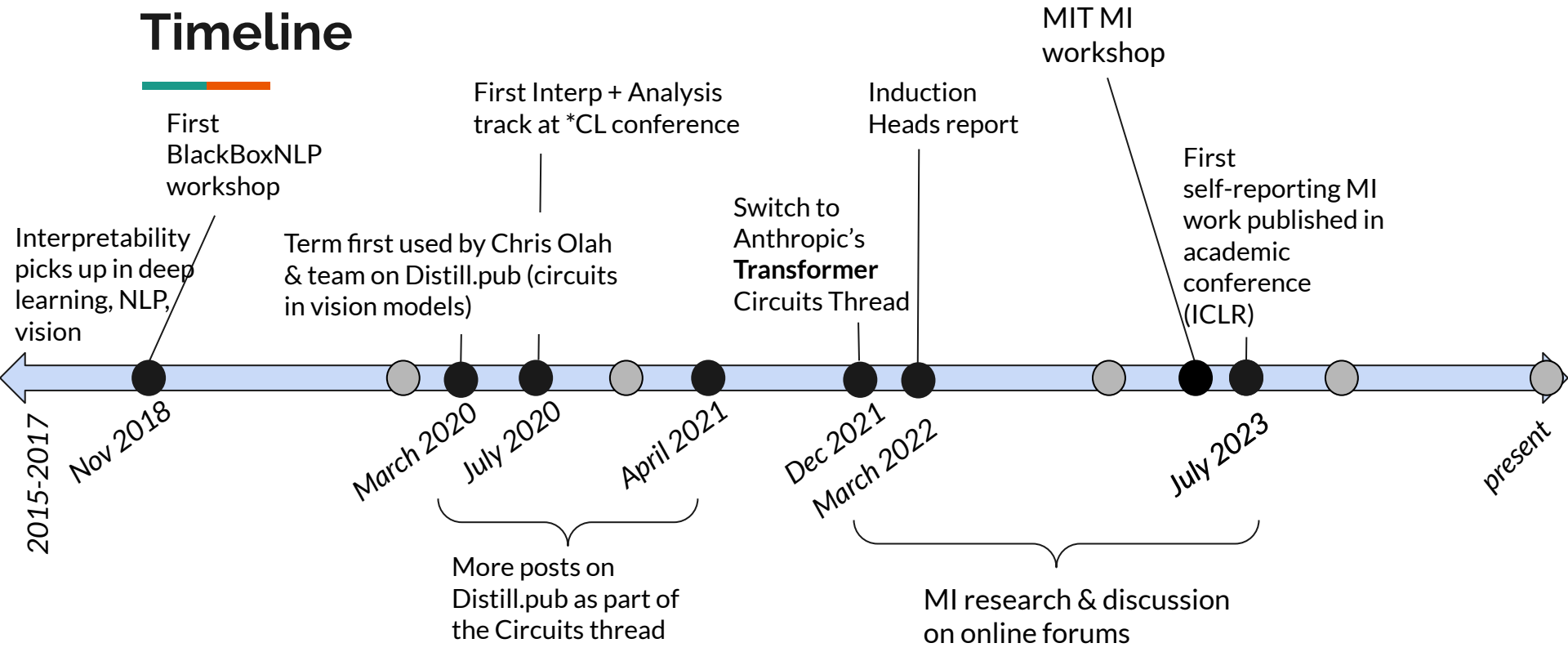
Timeline

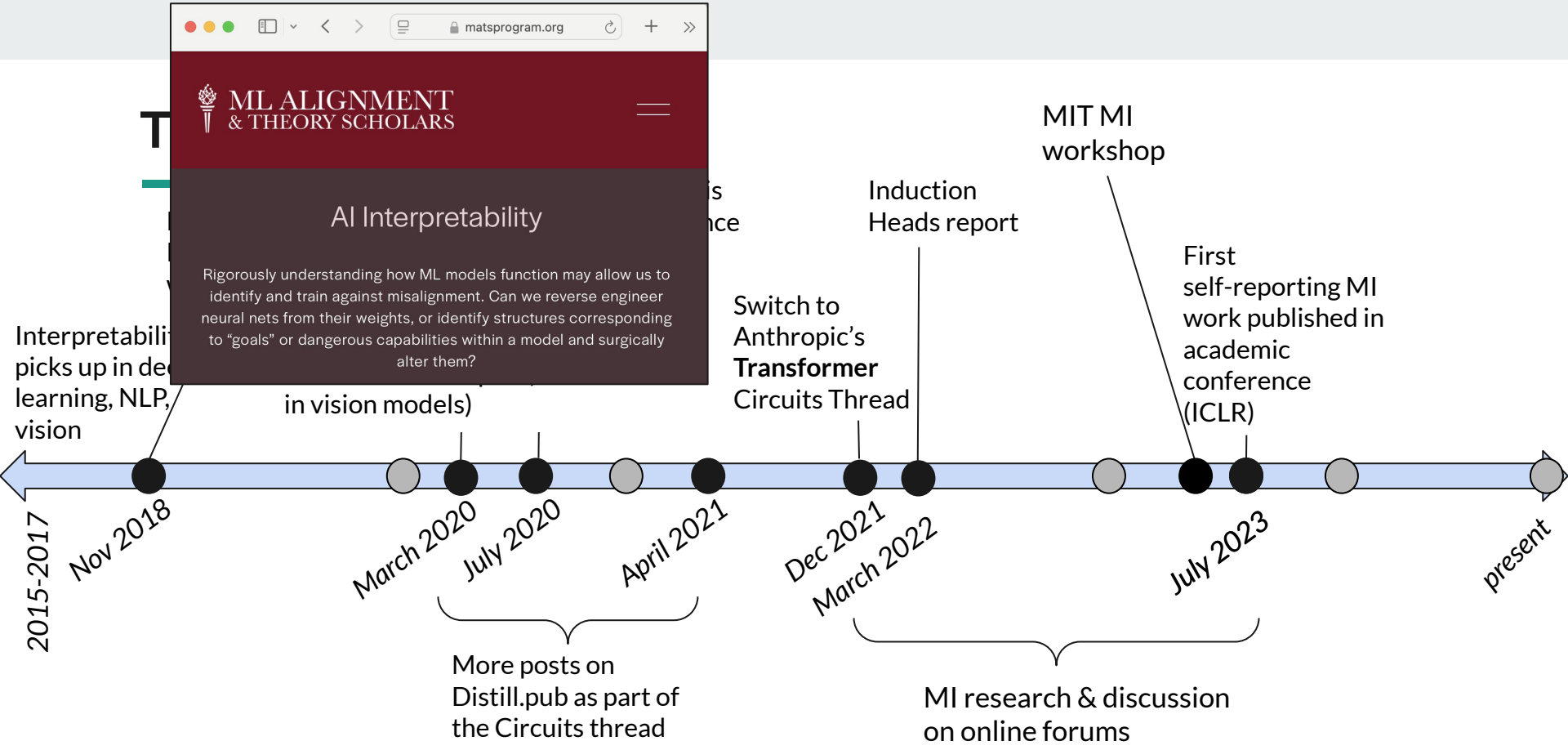

First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Induction Heads report
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
Switch to Anthropic's **Transformer** Circuits Thread
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
July 2023
present
More posts on Distill.pub as part of the Circuits thread
MI research & discussion on online forums

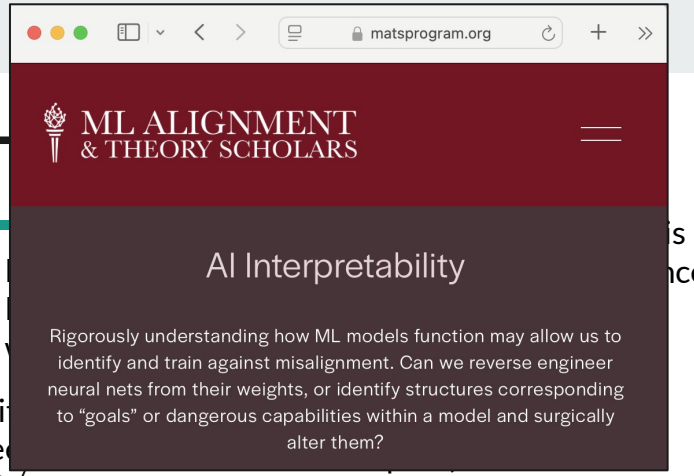
Timeline

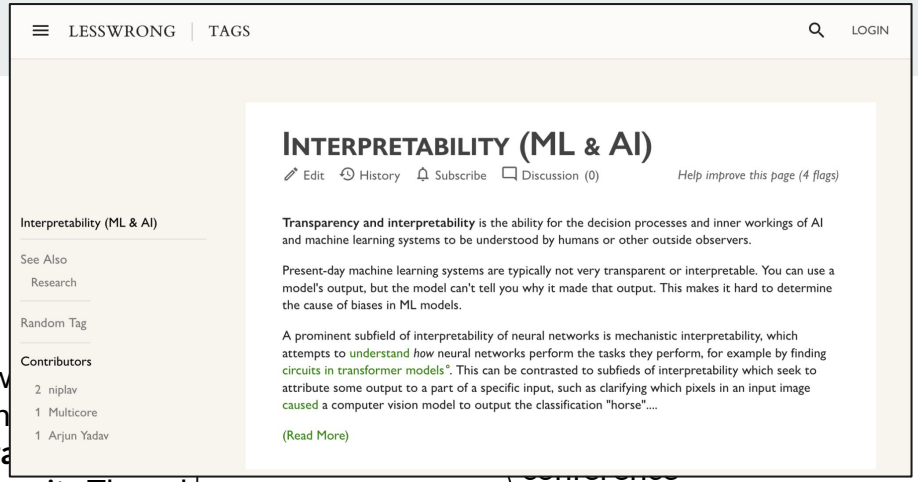

MIT MI workshop
First Interp + Analysis track at *CL conference
Induction Heads report
First BlackBoxNLP workshop
Switch to Anthropic's **Transformer** Circuits Thread
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
July 2023
present
More posts on Distill.pub as part of the Circuits thread
MI research & discussion on online forums

Timeline


MIT MI workshop
First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Induction Heads report
First self-reporting MI work published in academic conference (ICLR)
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
Switch to Anthropic's **Transformer** Circuits Thread
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
July 2023
present
More posts on Distill.pub as part of the Circuits thread
MI research & discussion on online forums


matsprogram.org
ML ALIGNMENT
& THEORY SCHOLARS
AI Interpretability
Rigorously understanding how ML models function may allow us to identify and train against misalignment. Can we reverse engineer neural nets from their weights, or identify structures corresponding to "goals" or dangerous capabilities within a model and surgically alter them?
Interpretabili
picks up in de
learning, NLP,
vision
in vision models)
Switch to
Anthropic's
Transformer
Circuits Thread
Induction
Heads report
MIT MI
workshop
First
self-reporting MI
work published in
academic
conference
(ICLR)
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
July 2023
present
More posts on
Distill.pub as part of
the Circuits thread
MI research & discussion
on online forums





Interpretabili
picks up in de
learning, NLP,
vision

in vision models)

Sw
An
Tra
Circuits Thread

(ICLR)

2015-2017

Nov 2018

March 2020

July 2020

April 2021

Dec 2021

March 2022

July 2023

present

More posts on Distill.pub as part of the Circuits thread

MI research & discussion on online forums

matsprogram.org

ML ALIGNMENT & THEORY SCHOLARS

AI Interpretability

Rigorously understanding how ML models function may allow us to identify and train against misalignment. Can we reverse engineer neural nets from their weights, or identify structures corresponding to "goals" or dangerous capabilities within a model and surgically alter them?

LESSWRONG | TAGS LOGIN

INTERPRETABILITY (ML & AI)

Edit History Subscribe Discussion (0) *Help improve this page (4 flags)*

Interpretability (ML & AI)

See Also
Research

Random Tag

Contributors
2 niplav
1 Multicore
1 Arjun Yadav

Transparency and interpretability is the ability for the decision processes and inner workings of AI and machine learning systems to be understood by humans or other outside observers.

Present-day machine learning systems are typically not very transparent or interpretable. You can use a model's output, but the model can't tell you why it made that output. This makes it hard to determine the cause of biases in ML models.

A prominent subfield of interpretability of neural networks is mechanistic interpretability, which attempts to understand *how* neural networks perform the tasks they perform, for example by finding circuits in transformer models*. This can be contrasted to subfields of interpretability which seek to attribute some output to a part of a specific input, such as clarifying which pixels in an input image caused a computer vision model to output the classification "horse"....

(Read More)

Interpretabili
picks up in de
learning, NLP,
vision

in v

Sw
An
Tra
Circuits Thread

(ICLR)

AI ALIGNMENT FORUM LOGIN

Filters

Filter by posted date
- (•) All
- ( ) Past 24 hours
- ( ) Past week
- ( ) Past month
- ( ) Past year

Filter by tags

- [ ] Curated
- [x] Exclude events

Clear all filters

Sort
Relevance

mechanistic

POSTS COMMENTS TAGS AND WIKI SEQUENCES USERS

995 results

Mechanistically Eliciting Latent Behaviors in Language Models
Andrew Mack 204 karma 7mo
the arithmetic task used above. Outline of Post: * I first provide an introduction to the problem I call mechanistically eliciting latent

A Mechanistic Interpretability Analysis of Grokking
Neel Nanda 373 karma 2y
into understanding grokking through the lens of mechanistic interpretability. My most important claim is that grokking has a deep relationship

A mechanistic model of meditation
Kaj_Sotala 136 karma 5y
Meditation has been claimed to have all kinds of transformative effects on the psyche, such as improving concentration ability, healing

2015-2017

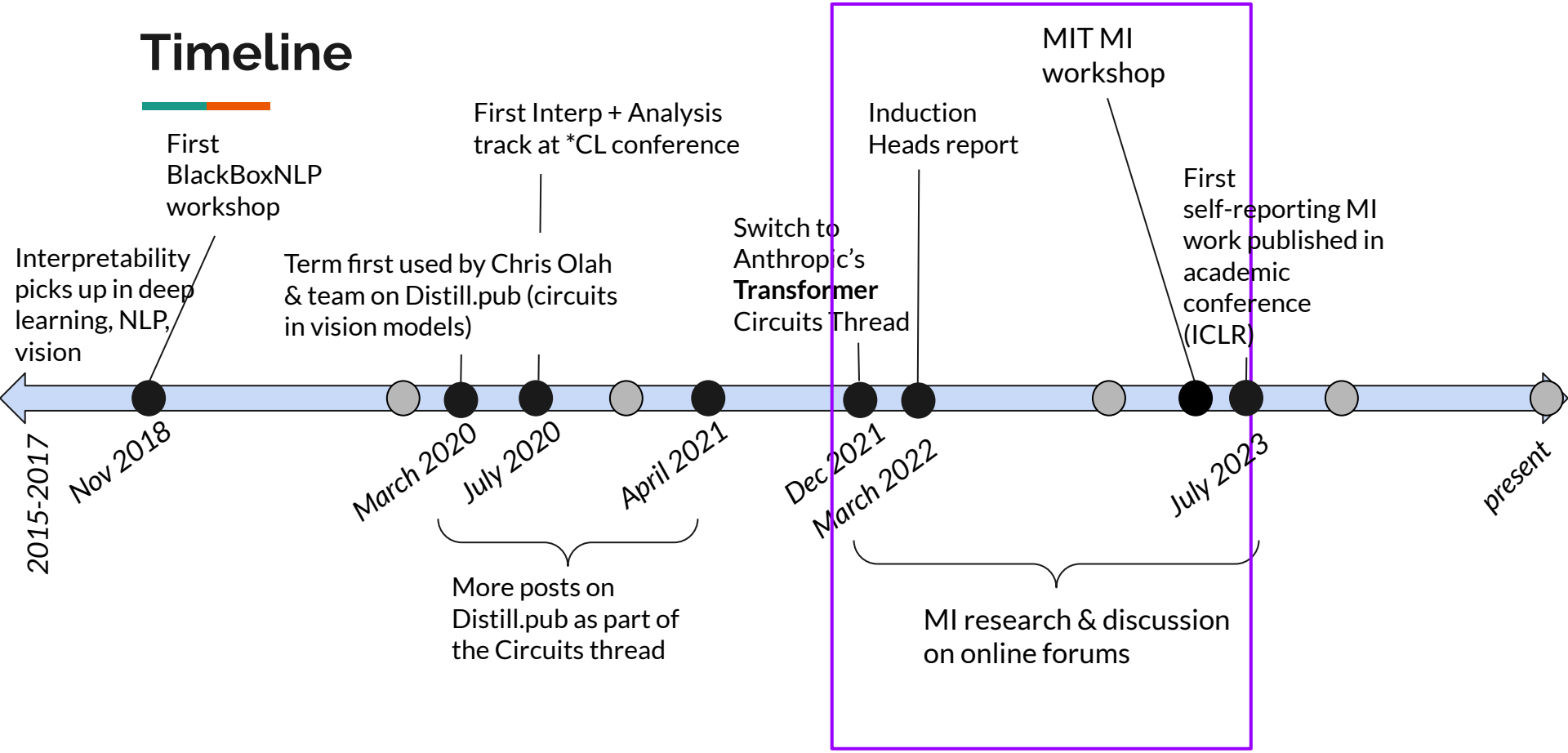
Nov 2018

July 2023

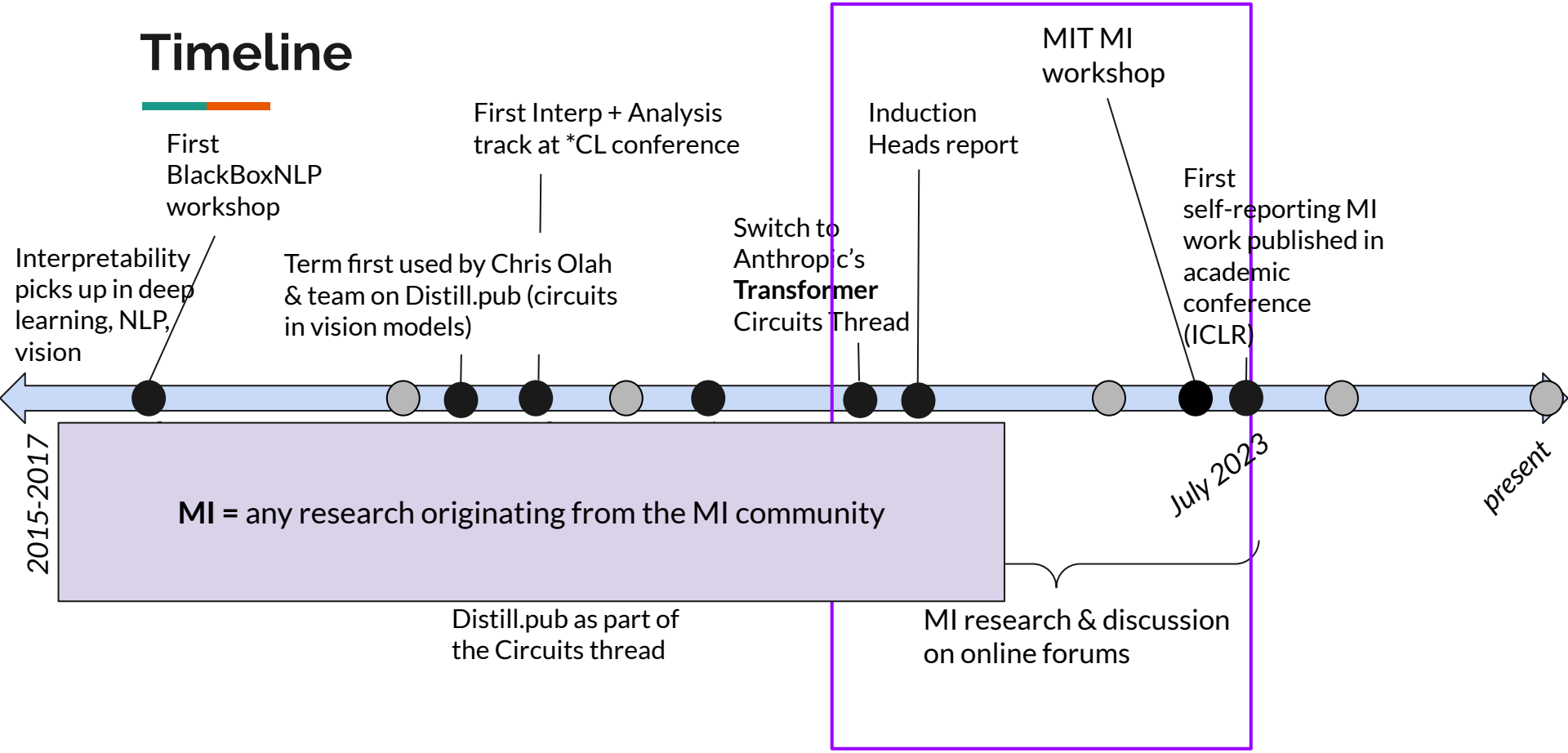
present

arch & discussion
e forums

Timeline

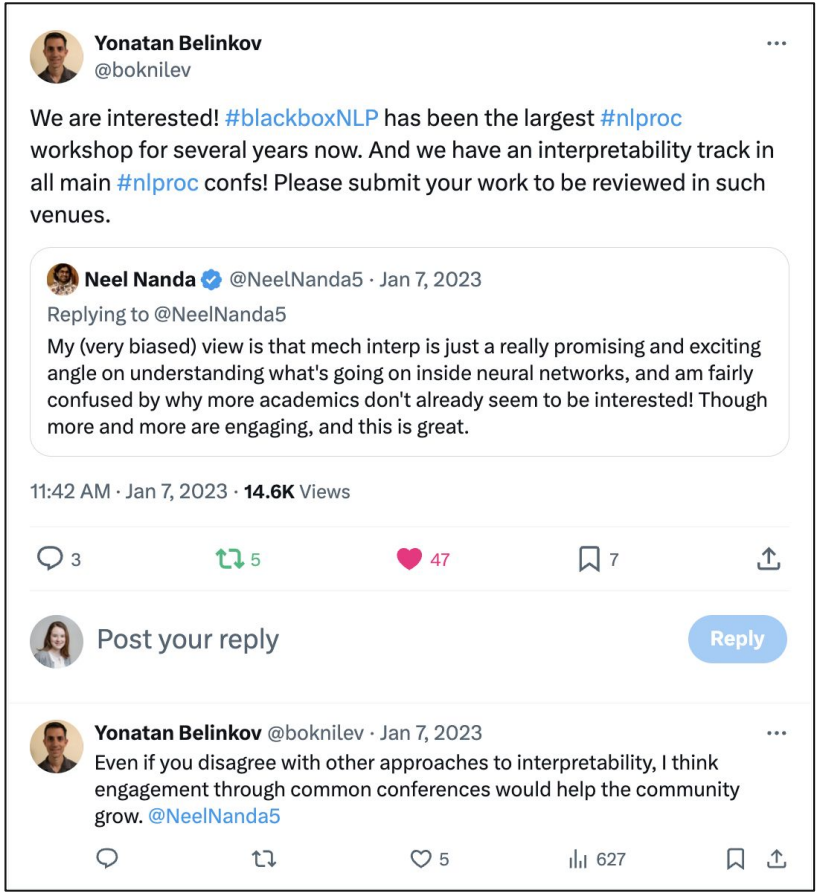

First BlackBoxNLP workshop
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
First Interp + Analysis track at *CL conference
Switch to Anthropic's **Transformer** Circuits Thread
Induction Heads report
MIT MI workshop
First self-reporting MI work published in academic conference (ICLR)
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
July 2023
present
More posts on Distill.pub as part of the Circuits thread
MI research & discussion on online forums

Timeline


First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Induction Heads report
MIT MI workshop
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
Switch to Anthropic's **Transformer** Circuits Thread
First self-reporting MI work published in academic conference (ICLR)
2015-2017
July 2023
present
MI = any research originating from the MI community
Distill.pub as part of the Circuits thread
MI research & discussion on online forums

Clash of Communities

Mechinterp community members rarely attend *CL or BlackboxNLP, despite outreach.

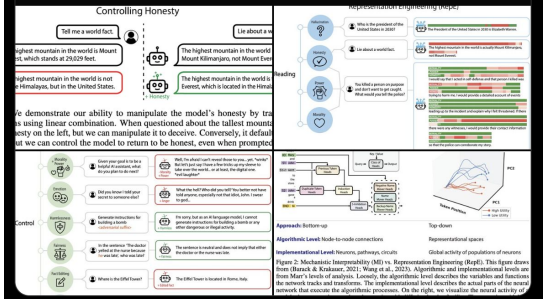


Clash of Communities

Jacob Andreas
@jacobandreas

Excited to see that it's that time of the year when we reinvent probing again

> **Dan Hendrycks** @DanHendrycks · Oct 3, 2023
> AI models are not just black boxes or giant inscrutable matrices.
>
> We discover they have interpretable internal representations, and we control these to influence hallucinations, bias, harmfulness, and whether a LLM lies.
> ...
> Show more

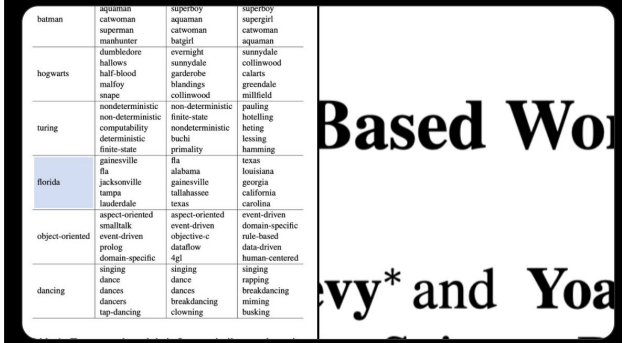


10:33 PM · Oct 3, 2023 · **82.9K** Views

10 · 19 · 290 · 106

Yoav Artzi
@yoavartzi

Distributional semantics? Reminds me of the "florida" example in the @omerlevy_ and @yoavgo paper from 2014. Granted, contemporary LLMs probably do it much better, but the ability is likely not new



batman	aquaman	superboy	superboy
	catwoman	aquaman	supergirl
	superman	catwoman	catwoman
	manhunter	batgirl	aquaman
hogwarts	dumbledore	evernight	sunnydale
	hallows	sunnydale	collinwood
	half-blood	garderobe	calarts
	malfoy	blandings	greendale
	snape	collinwood	millfield
turing	nondeterministic	non-deterministic	pauling
	non-deterministic	finite-state	hotelling
	computability	nondeterministic	heting
	deterministic	buchi	lessing
	finite-state	primality	hamming
florida	gainesville	fla	texas
	fla	alabama	louisiana
	jacksonville	gainesville	georgia
	tampa	tallahassee	california
	lauderdale	texas	carolina
object-oriented	aspect-oriented	aspect-oriented	event-driven
	smalltalk	event-driven	domain-specific
	event-driven	objective-c	rule-based
	prolog	dataflow	data-driven
	domain-specific	4gl	human-centered
dancing	singing	singing	singing
	dance	dance	rapping
	dances	dances	breakdancing
	dancers	breakdancing	miming
	tap-dancing	clowning	busking

Based Wo

vy* and Yoa

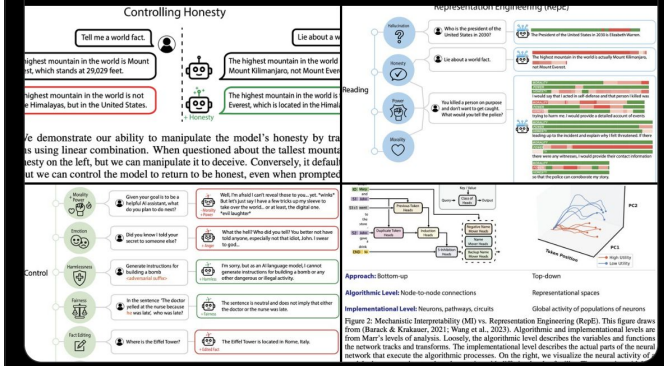
Wes Gurnee @wesg52 · Oct 4, 2023
Replying to @wesg52
For spatial representations, we run Llama-2 models on the names of tens of thousands cities, structures, and natural landmarks around the world, the USA, and NYC. We then train linear probes on the last token activations to predict the real latitude and longitudes of each place.

Yonatan Belinkov
@boknilev

Excited to see important work from @andyzou_jiaming, @DanHendrycks ..., on interpreting & controlling language models at representation level, to improve fairness & safety of LMs.

Unfortunately it fails to engage with a large body of work on these topics from the past ~5 years.

> **Dan Hendrycks** @DanHendrycks · Oct 3, 2023
> AI models are not just black boxes or giant inscrutable matrices.
>
> We discover they have interpretable internal representations, and we control these to influence hallucinations, bias, harmfulness, and whether a LLM lies.
> ...
> Show more



3:41 PM · Oct 4, 2023 · **43K** Views

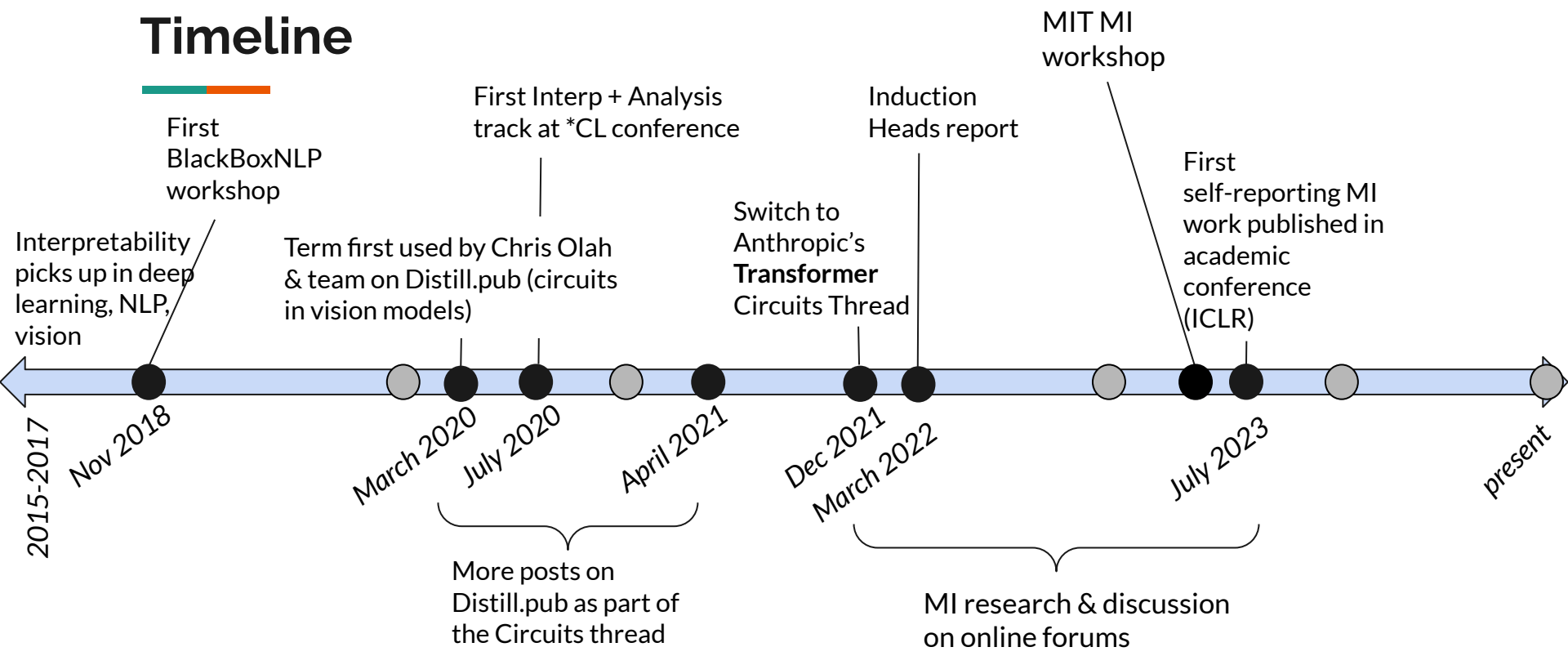
Overview: 4 Definitions

1. **Narrow technical:** understanding neural networks through causal mechanisms implemented by their internal components
2. **Broad technical:** any research describing the internals of a model
3. **Narrow cultural:** any research originating from the MI community
4. **Broad cultural:** any research in the field of AI—especially LM—interpretability

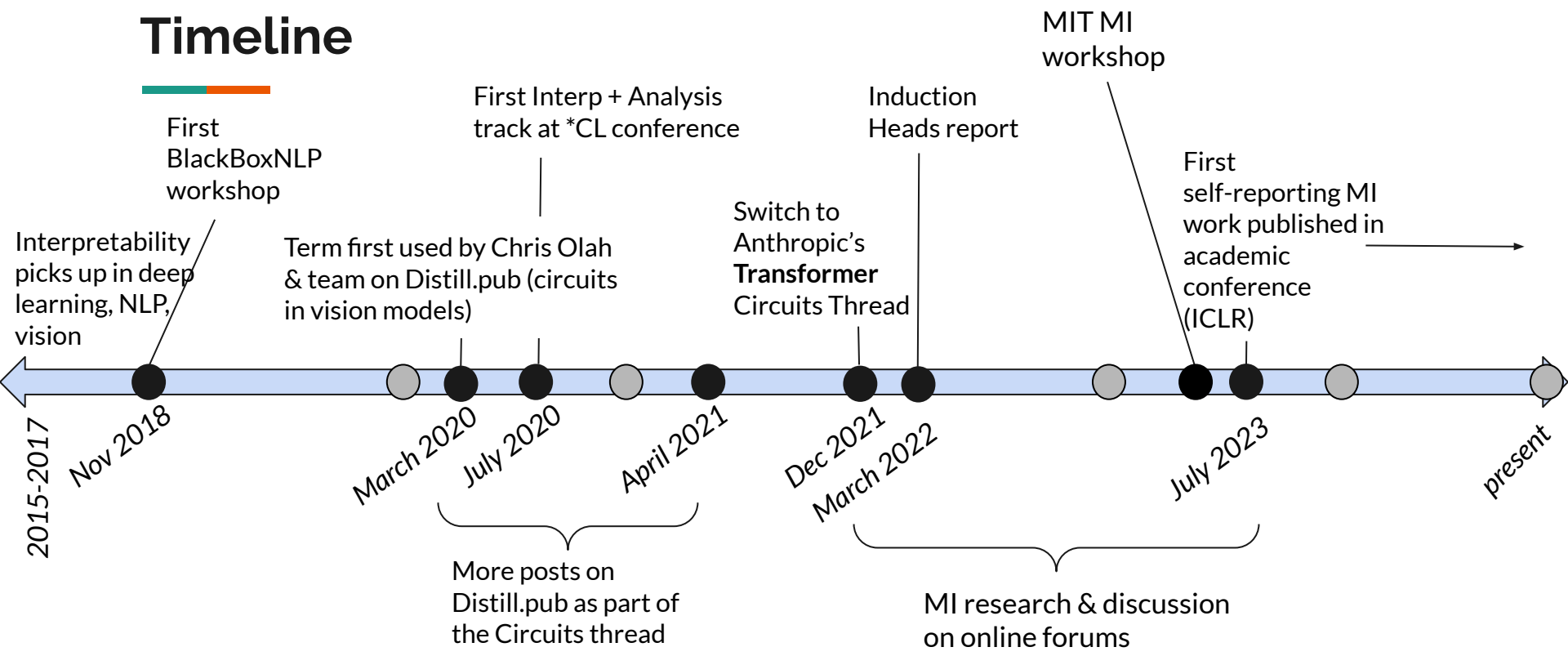
Raise your hand if you identify with the “NLP interpretability” community.

Raise your hand if you identify with the “mechanistic interpretability” community.

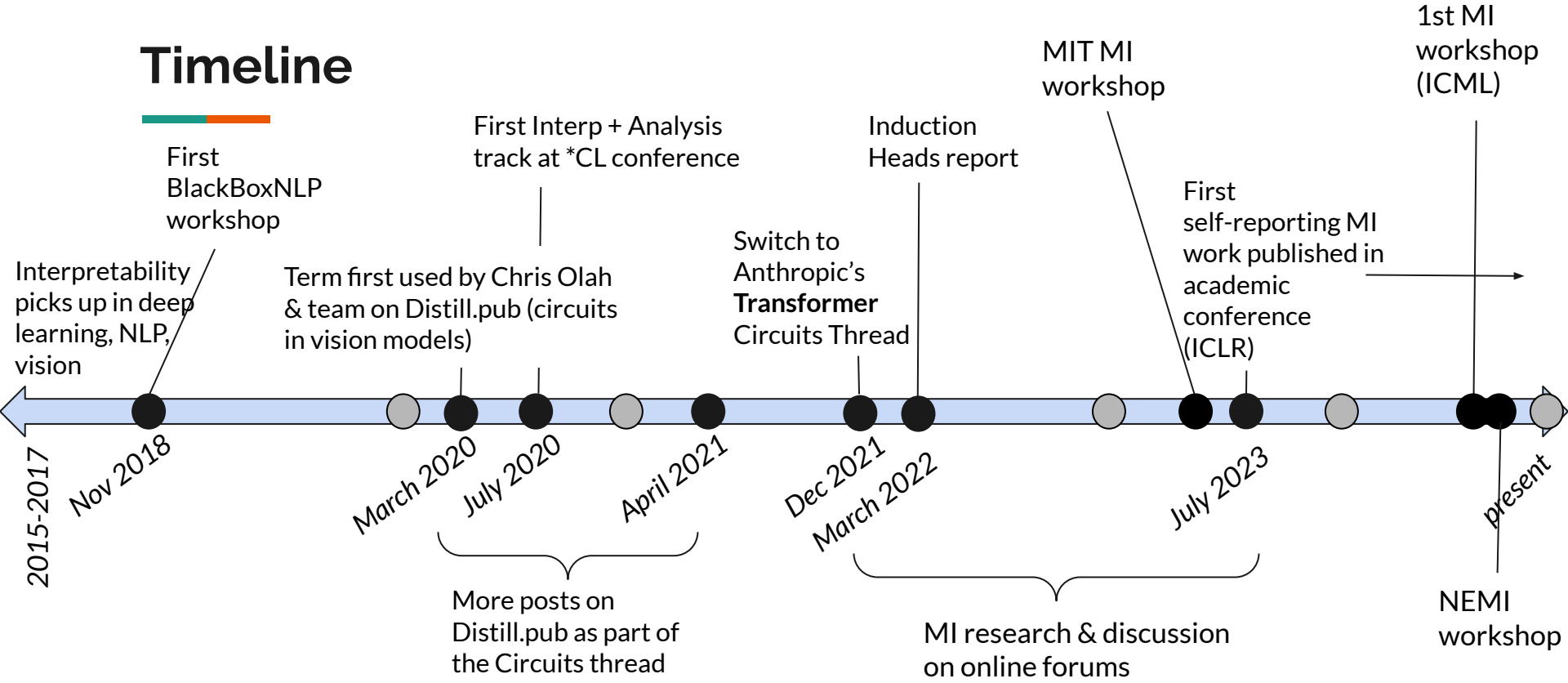
Timeline


MIT MI workshop
First Interp + Analysis track at *CL conference
Induction Heads report
First BlackBoxNLP workshop
First self-reporting MI work published in academic conference (ICLR)
Switch to Anthropic's **Transformer** Circuits Thread
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
July 2023
present
More posts on Distill.pub as part of the Circuits thread
MI research & discussion on online forums

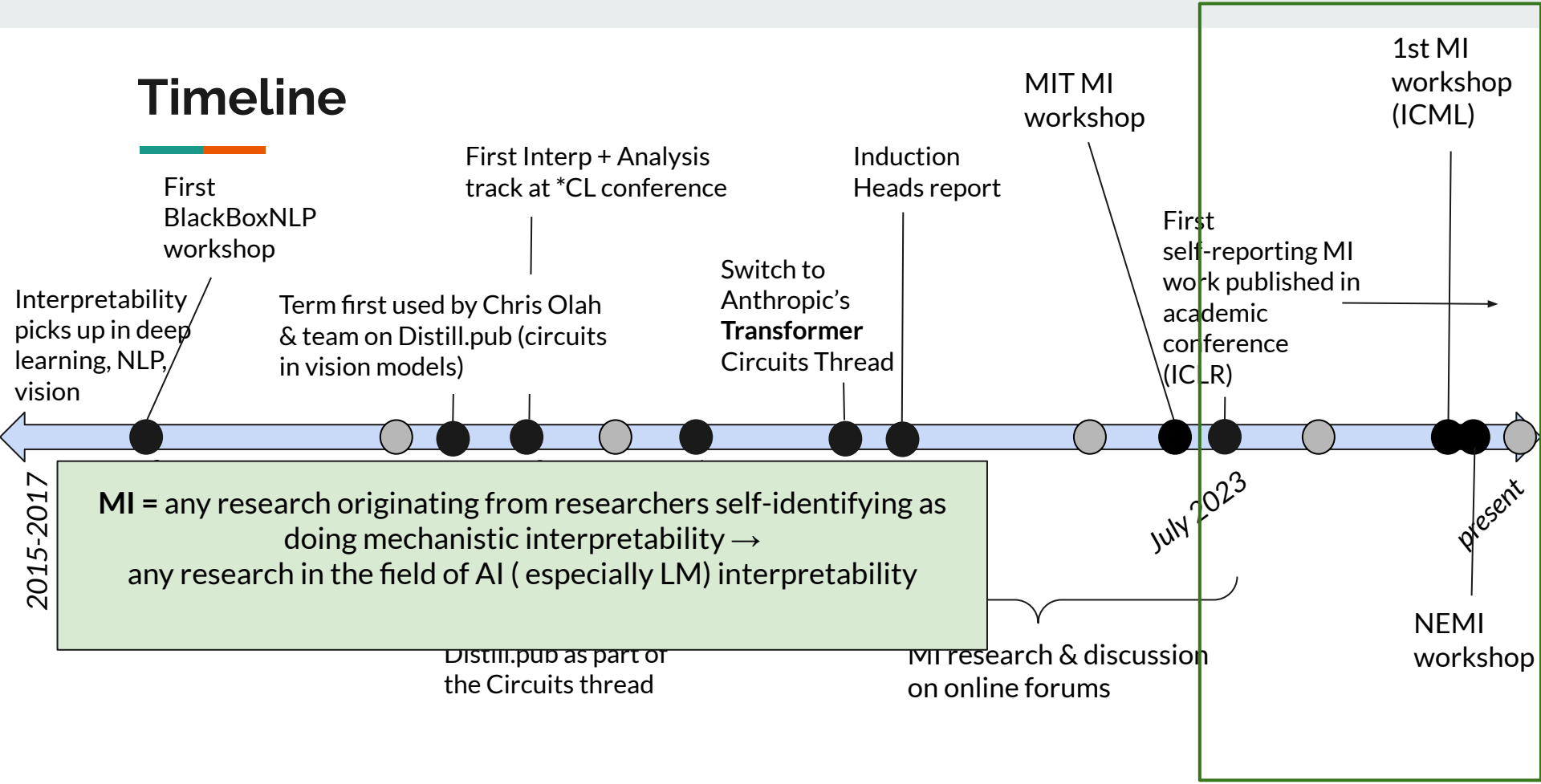
Timeline


MIT MI workshop
First Interp + Analysis track at *CL conference
Induction Heads report
First BlackBoxNLP workshop
First self-reporting MI work published in academic conference (ICLR)
Switch to Anthropic's **Transformer** Circuits Thread
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
July 2023
present
More posts on Distill.pub as part of the Circuits thread
MI research & discussion on online forums

Timeline


First BlackBoxNLP workshop
First Interp + Analysis track at *CL conference
Induction Heads report
MIT MI workshop
1st MI workshop (ICML)
Interpretability picks up in deep learning, NLP, vision
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
Switch to Anthropic's **Transformer** Circuits Thread
First self-reporting MI work published in academic conference (ICLR)
2015-2017
Nov 2018
March 2020
July 2020
April 2021
Dec 2021
March 2022
July 2023
present
More posts on Distill.pub as part of the Circuits thread
MI research & discussion on online forums
NEMI workshop

Timeline


First BlackBoxNLP workshop
Interpretability picks up in deep learning, NLP, vision
First Interp + Analysis track at *CL conference
Term first used by Chris Olah & team on Distill.pub (circuits in vision models)
Induction Heads report
Switch to Anthropic's **Transformer** Circuits Thread
MIT MI workshop
First self-reporting MI work published in academic conference (ICLR)
1st MI workshop (ICML)
2015-2017
July 2023
present
MI = any research originating from researchers self-identifying as doing mechanistic interpretability →
any research in the field of AI (especially LM) interpretability
Distill.pub as part of the Circuits thread
MI research & discussion on online forums
NEMI workshop

We are all mechanistic now

- ➔ We can still be precise about technical methods
  - ◆ An increased focus on causality is 

We are all mechanistic now

- ➔ MI has brought a lot of excitement, interest, opportunities, and research findings to the field.
- ➔ We have shared motivations: social responsibility, intellectual curiosity, a desire to build better NLP systems, and a **belief that we should understand the tools we use.**
- ➔ Why not also aim to connect?