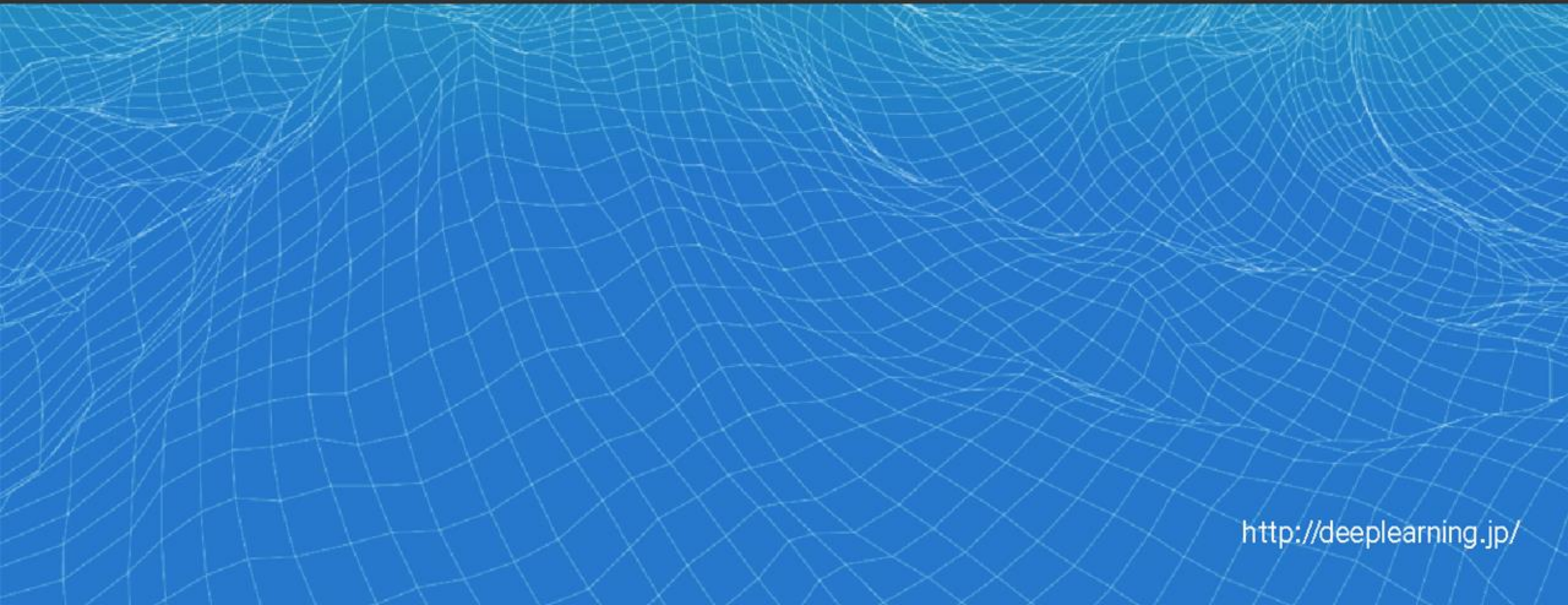
DEEP LEARNING JP
[DL Papers]

Mechanistic Interpretability for AI Safety: A Review

Presenter: Yoshimasa Tawatsuji, Matsuo-Iwasawa Lab



http://deeplearning.jp/

書誌情報

- Mechanistic Interpretability for AI Safety: A Review
  - 著者：Leonard Bereska, Efstratios Gavves
  - 所属：University of Amsterdam

- 概要
  - AIの内部構造を解明し、安全性や価値の整合性を確保するために重要な「機械論的解釈可能性」を概観。ニューラルネットワークが学習した計算メカニズムを人間が理解可能な形に解析し、詳細かつ因果的な理解を目指す。
  - 本論文では、AI安全性におけるこの手法の意義、課題を調査し、将来的な方向性について議論。

全体の章構成

- Introduction
- Interpretability Paradigms from the Outside In
- Core Concepts And Assumptions
  - Defining Features as Representational Primitives
  - Nature of Features: From Monosemantic Neurons to Non-Linear Representations
  - Circuits as Computational Primitives and Motifs as Universal Circuit Patterns
  - Emergence of World Models and Simulated Agents
- Core Methods
  - Taxonomy of Mechanistic Interpretability Methods
  - Observation
  - Intervention
    - Active Patching
    - Causal Abstraction
    - Hypothesis Testing
  - Integrating Observation and Intervention
- Current Research
  - Intrinsic Interpretability
  - Developmental Interpretability
  - Post-Hoc Interpretability
  - Automation: Scaling Post-Hoc Interpretability
- Relevance to AI Safety
- Challenges
  - Research Issues
  - Technical Limitations
- Future Directions
  - Clarifying Concepts
  - Setting Standards
  - Scaling Techniques
  - Expanding Scope

Introduction

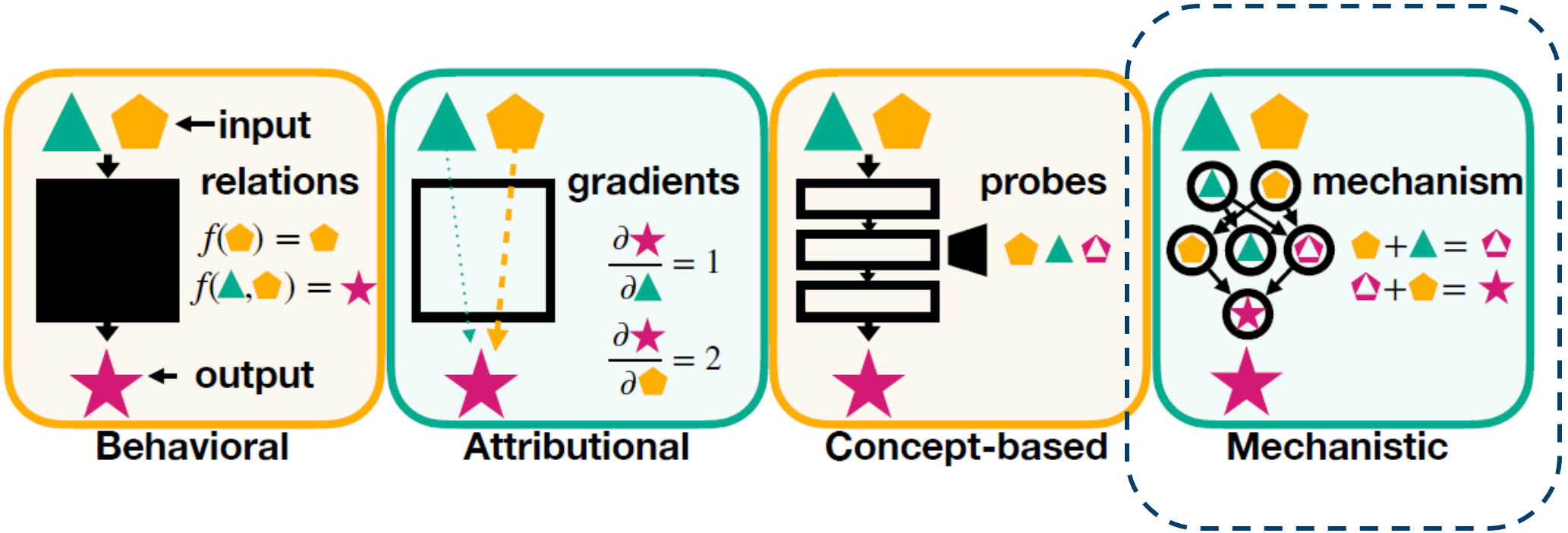
- AIシステムの価値整合性や安全性の確保が重要

- 「機械論的解釈可能性（Mechanistic Interpretability）」を中心とした、解釈可能性研究の新たなアプローチとして位置づけ
  - ニューラルネットワークの内部計算メカニズムを詳細に解析し、人間が理解可能な形での解釈を可能

- AIシステムがブラックボックスではなく、因果的かつ詳細な分析に基づく透明性を持つことが重要

Interpretability Paradigms from the Outside In

- AIシステムの解釈手法の「外部から内部へ」という視点で分類



Core Concepts And Assumptions

- Defining Features as Representational Primitives
- Nature of Features: From Monosemantic Neurons to Non-Linear Representations
- Circuits as Computational Primitives and Motifs as Universal Circuit Patterns
- Emergence of World Models and Simulated Agents

Defining Features as Representational Primitives

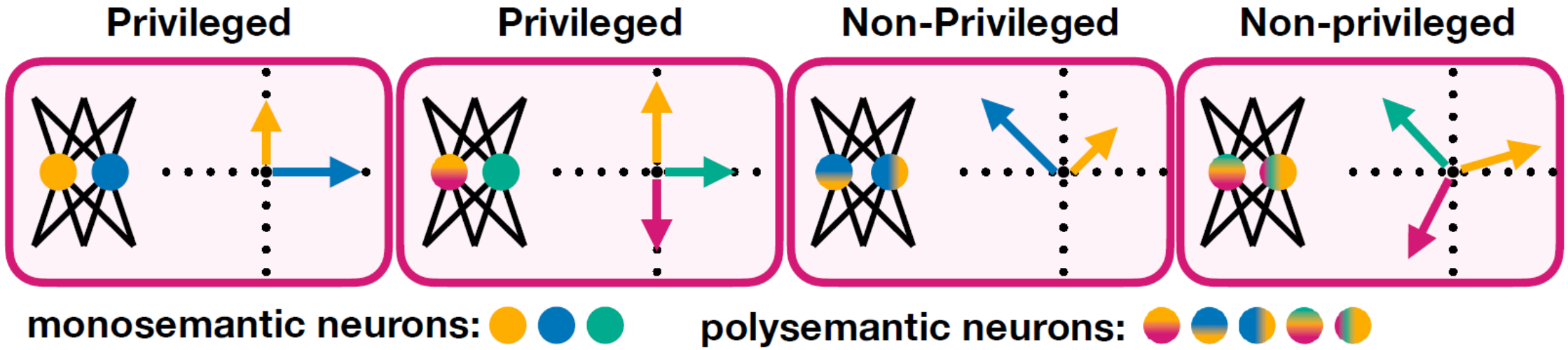
定義：特徴（Feature）

ニューラルネットワークの表現の基本単位であり，これ以上単純な要素へ機能分解が不能なもの

- 自然な抽象物（natural abstractions）を構成するコンパクトな表現
- 表現素子（representation atoms）
- 人間の解釈可能性を超えた特徴
  - 機械論的解釈可能性では、特徴が人間の概念から逸脱している場合でも、学習した実際の表現を明らかにすることを目的としている
  - 人間の概念と一致しているかどうかに関係なく、機能を独立した（不可分な）モデルコンポーネントとして定義することはより包括的なアプローチ

Nature of Features: From Monosemantic Neurons to Non-Linear Representations

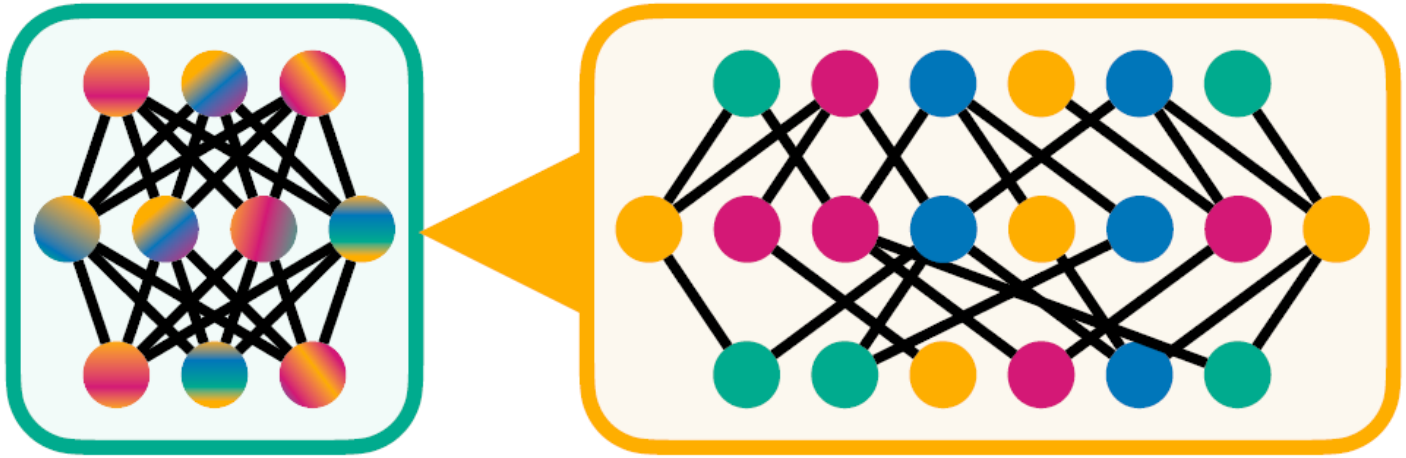
- 単一ニューロンの特徴は「Monosemantic」か「Polysemantic」か
  - Transformer model は経験上「Polysemantic」
  - 多義性はニューロンの表現素子としての性質と矛盾し、解釈可能性を困難にする要因



Nature of Features: From Monosemantic Neurons to Non-Linear Representations

重ね合わせ仮説（Superposition Hypothesis）

ニューラル ネットワークは、ニューロンの重複した組み合わせで特徴をエンコードすることで、ニューロンの数よりも多くの特徴を表現する。



Nature of Features: From Monosemantic Neurons to Non-Linear Representations

再定義：Feature

ニューロン数が制限要因でない場合、ネットワークが理想的には個々のニューロンに割り当てる要素 (Bricken et al., 2023)。言い換えれば、特徴は、十分な容量を持つより大規模でスパースなネットワークが個々のニューロンで表現を学習する、disentangled concepts（解きほぐされた概念）に対応

- では、個々のNeuronで特徴がエンコードされていないなら、特徴は一体どこにあるのか？
  - ☞ **線型表現仮説**：ニューロンの線型結合（表現空間の「方向」（direction））

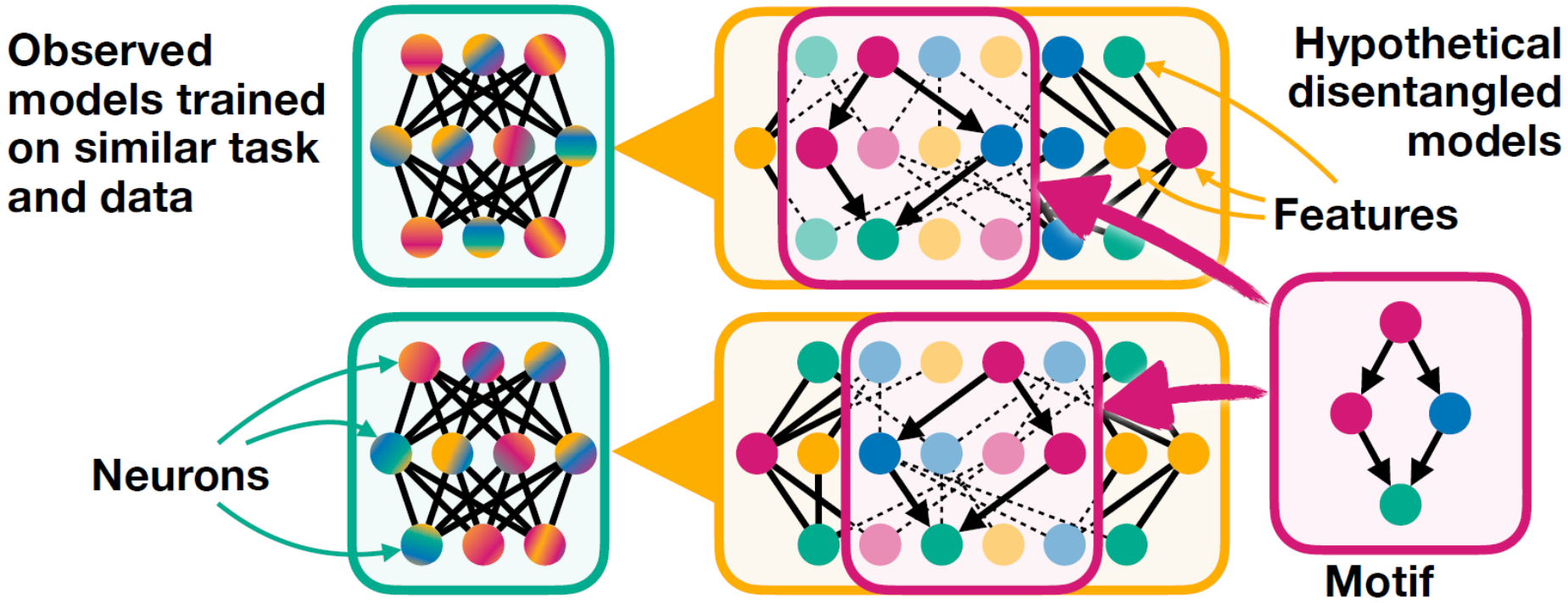
Circuits as Computational Primitives and Motifs as Universal Circuit Patterns

定義：Circuit（回路）

ネットワークのサブグラフであり、特徴とそれらを接続する重みで構成される

定義：Motif

ネットワーク内で繰り返されるパターンであり、様々なモデルやタスクにわたって出現する特徴あるいは回路のいずれかを含むもの



Circuits as Computational Primitives and Motifs as Universal Circuit Patterns

- ニューラルネットワークの特徴と回路の収束に関する２つの普遍性仮説

弱い普遍性（Weak Universality）

ニューラルネットワークが特定のタスクを解決する方法を学習する方法には、基本原則がある。モデルは一般に、共通の基本原則に準拠した類似のソリューションに収束する。ただし、これらの原則を実装する特定の特徴と回路は、ハイパーパラメータ、ランダムシード、アーキテクチャの選択などの要因に基づいて、モデルごとに異なる。

強い普遍性（Strong Universality）

同様のタスクとデータ分布でトレーニングされ、同様の手法を使用しているすべてのニューラルネットワークモデルでは、同じコアとなる特徴と回路が普遍的かつ一貫して出現する。これは、ニューラルネットワークが学習時に本質的に引き寄せられる一連の基本的な計算モチーフを反映している。

Emergence of World Models and Simulated Agents

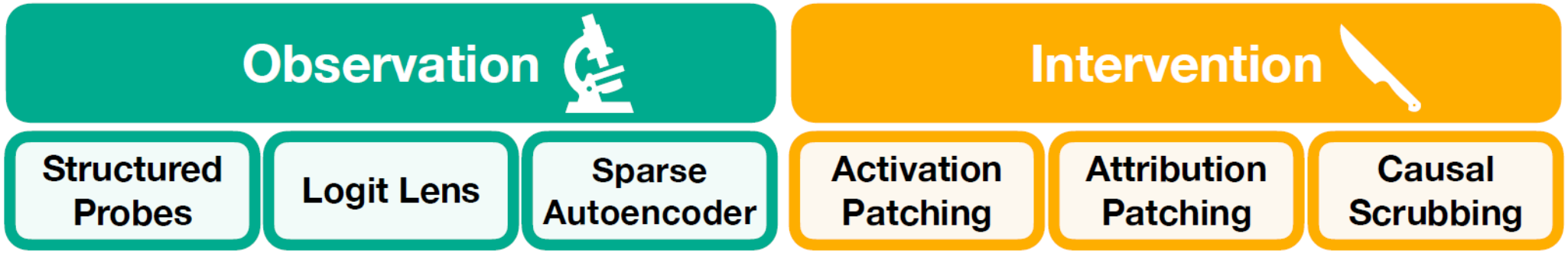
- 内部世界モデル（Internal World Model）
  - ニューラルネットワーク内で形成される環境の内部因果モデル

- 「確率的オウム」としてのLLM：
  - 相関関係は学習できるが介入データにアクセスできず、世界の因果モデルを開発する能力が欠けている

- アクティブ推論とシミュレーション仮説
  - 予測誤差を最小化する目標は、複雑な世界表現を形成するのに十分な条件とされ、LLMもこれに従って言語や世界モデルを構築しうる
  - 十分に訓練された予測モデルがデータ生成の因果過程を模倣するようになり、GPTのようなモデルが自然に内部世界モデルを発展しうる

Core Methods

- Taxonomy of Mechanistic Interpretability Methods
- Observation
- Intervention
  - Active Patching
  - Causal Abstraction
  - Hypothesis Testing
- Integrating Observation and Intervention



Taxonomy of Mechanistic Interpretability Methods

Table 1: Taxonomy of Mechanistic Interpretability Methods

Method	Causal Nature	Phase	Locality	Comprehensiveness	Key Examples
Feature Visualization	Observation	Post-hoc	Local	Partial	Zeiler & Fergus (2014) Zimmermann et al. (2021)
Exemplar methods	Observation	Post-hoc	Local	Partial	Grosse et al. (2023) Garde et al. (2023)
Probing Techniques	Observation	Post-hoc	Both	Both	McGrath et al. (2022) Gurnee et al. (2023)
Structured Probes	Observation	Post-hoc	Both	Both	Burns et al. (2023)
Logit Lens Variants	Observation	Post-hoc	Global	Partial	nostalgebraist (2020) Belrose et al. (2023)
Sparse Autoencoders	Observation	Post-hoc	Both	Comprehensive	Cunningham et al. (2024) Bricken et al. (2023)
Activation Patching	Intervention	Post-hoc	Local	Partial	Meng et al. (2022a) Wang et al. (2023)
Path Patching	Intervention	Post-hoc	Both	Both	Goldowsky-Dill et al. (2023)
Causal Abstraction	Intervention	Post-hoc	Global	Comprehensive	Geiger et al. (2023a) Geiger et al. (2023b) Wu et al. (2023a)
Hypothesis Testing	Intervention	Post-hoc	Global	Comprehensive	Chan et al. (2022) Jenner et al. (2023)
Intrinsic Methods	–	Pre/During	Global	Comprehensive	Elhage et al. (2022a) Liu et al. (2023a)

Observation / Innervation

- 観察的手法：モデル内部の特徴や表現を分析する方法
  - 特徴の可視化
  - プロービング
  - Logit Lens
  - スパース辞書学習（重複した特徴の分解）

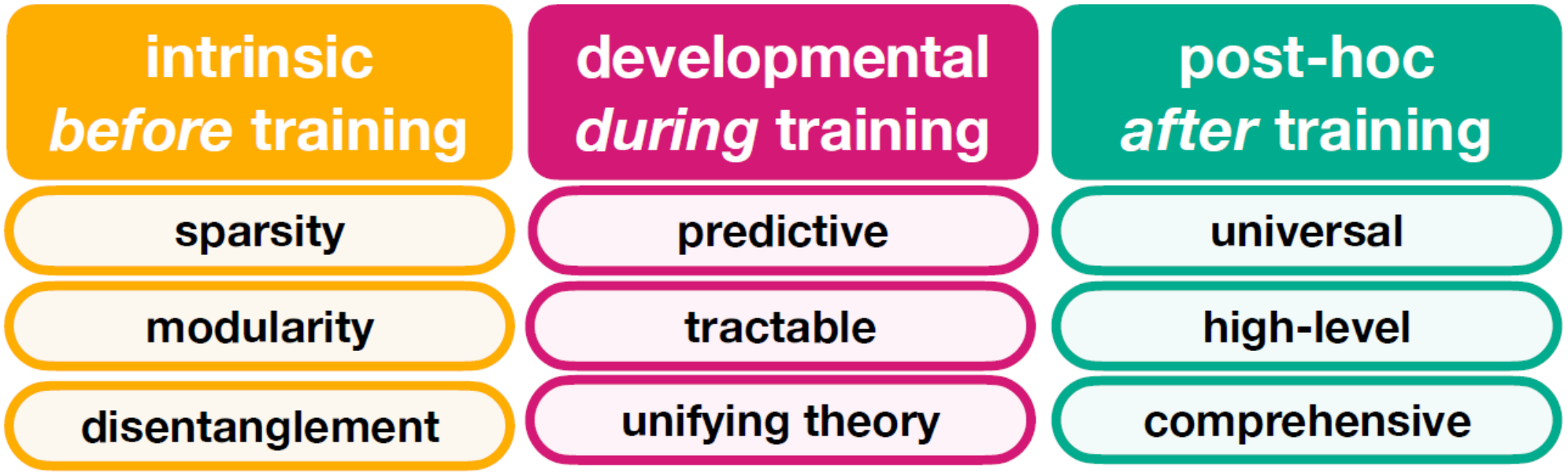
- 介入的手法：
  - Active Patching
  - Causal Abstraction
  - Hypothesis Testing

Integrating Observation and Intervention

- 観察と介入の統合的活用による包括的なモデルの理解方法
  - スパースオートエンコーダやロジットレンズの組み合わせ
  - 層ごとの予測形成や因果関係の特定に活用する方法

Current Research

- Intrinsic Interpretability
- Developmental Interpretability
- Post-Hoc Interpretability
- Automation: Scaling Post-Hoc Interpretability



Intrinsic Interpretability

- 内的解釈可能性の手法：
  - スパース性、モジュラリティ、モノセマンティックな構造の確保を目的としたアーキテクチャ設計や訓練プロセスの制約
  - モデルの解釈を容易にし、パフォーマンスを損なわずに理解を深める方法

Developmental Interpretability

- 開発的解釈可能性：
  - 学習過程における特徴や回路の形成を分析し、段階的な変化に対応する内部構造の出現を捉える。
  - 学習中の重要なフェーズ変化を予測または制御する可能性が示されており、AIモデルの安全性や信頼性に対する新たな知見を提供することが期待

Post-Hoc Interpretability

- 事後的解釈可能性
  - 目的：学習後のモデル解析に焦点。特定の振舞いや決定プロセスの理解
  - グローバル／ローカル解釈可能性
  - 包括的／部分的解釈可能性

- 普遍性
  - 様々なモデルやタスクに適用可能な一般的原則を明らかにする
  - 異なるモデルが同じ回路や特徴を必ずしも学習するわけではなく回路の形成や開発順序にはモデルやタスクによってばらつきがある可能性
- 高レベルの概念やタスクの符号化
  - 内部表現に介入して、モデルが高レベルの概念やタスクをどのように符号化しているのかに関する研究も

Automation: Scaling Post-Hoc Interpretability

- 事後的解釈可能性のワークフローの自動化に関するスケーラブルなアプローチ
  - 大規模モデルにおける解析の負担を軽減するための方法
  - 自動的に重要な回路を検出する技術

- モデルを解釈可能な回路へ分解する技術
  - 自動回路検出技術：特定のタスクに対するモデルの動作を支える重要な計算サブ回路またはコンポーネントを特定
  - サブタスクを解決するコンポーネントが満たすべき要件からサブコンポーネントを抽出する技術
- 抽出された回路の解釈技術
  - 大規模な言語モデル自体を解釈ツールとして使用する例など

Relevance to AI Safety

- 機械的解釈可能性がAI安全性にどのように関与し得るかについて説明
  - 特に、AIモデルの内部メカニズムを因果的に解釈することが、安全で信頼性の高いAIシステムの設計にどのように貢献するか
  - 機械的解釈可能性は、モデルの動作や判断基準の透明性を高め、リスクやバイアスの検出、意図せぬ能力の増大による安全性への懸念に対応するための重要なツール

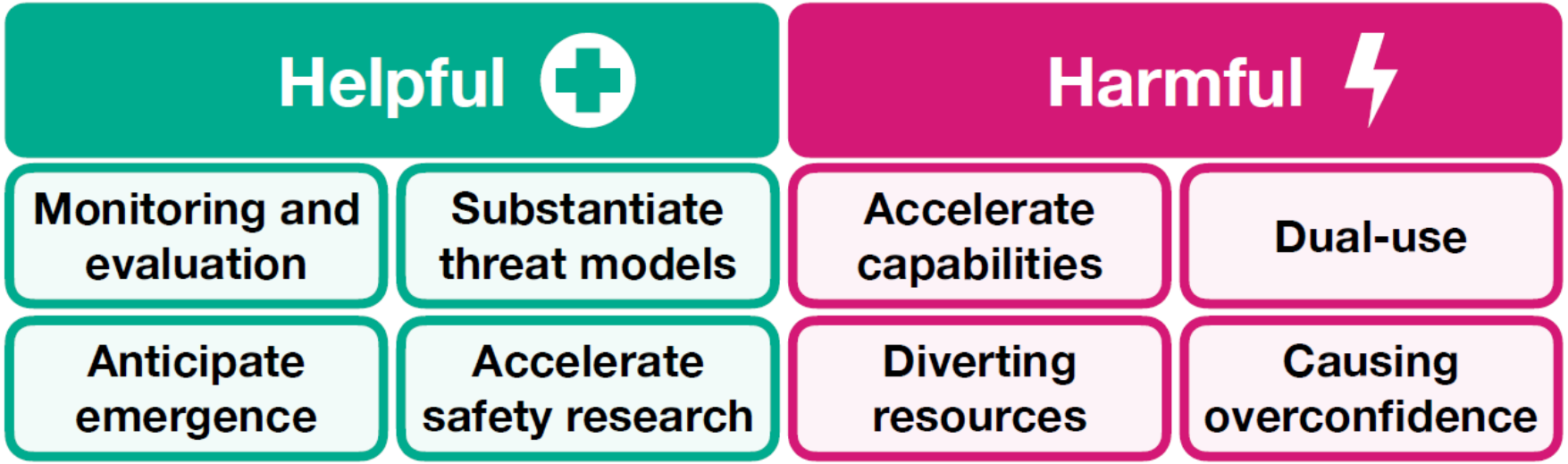


Figure 11: Potential benefits and risks of mechanistic interpretability for AI safety.

How could interpretability promote AI safety?

- 機械論的解釈可能性の意義
  - 解釈可能性ツールは、AIの情報処理や意思決定の理解を助け、モデルの評価を通じたフィードバックの強化や新たな能力の出現の予測を可能にする
  - リスクモデルに具体的な証拠を提供し、AIコミュニティでの安全性に対する認識を高める
- AIリスクへの対応
  - 悪用防止や競争圧力の軽減、AIアライメント（AIが意図された目標を追求すること）支援などに貢献
  - トレーニングの各段階で安全フィルターを提供し、モデルの誠実性や欺瞞行動を監視するための基盤を提供
- LLMの内部世界モデル
  - 人間の価値観を内部的に表現することでアライメントが容易になる
  - 特定の目標やエージェンシーを排除することでより安全とされる

How could interpretability promote AI safety?

- 強化学習とエージェンシーの問題
  - 強化学習はエージェントを生成しやすいため、AIシステムがエージェントとして振る舞うことのリスクが存在
  - 予測重視のモデルはエージェンシーを持たずともエージェント的な行動を模倣するため、内部にエージェンシーの要素があるかを慎重に監視する必要有
- AIアライメントへの統合
  - 人間の価値観を内部表現として認識し、それに基づいた目標を追求することが可能であれば、解釈可能性はそのままアライメント戦略となりうる
- リスクのスペクトラム
  - AIの安全性リスク：モデル中心のリスクと社会的なリスク
    - 機械論的解釈可能性はモデル中心リスクに対処する上で有効
    - 社会的リスク（経済構造の変化や進化的ダイナミクス）には他の分野の研究が補完的に必要

How could mechanistic insight be harmful?

- 解釈可能性研究のリスク
  - 機械論的解釈可能性の研究は、AI能力を加速させる可能性があり、これが人間の価値観と乖離した強力なAIの開発につながるリスク
  - リスク軽減のための選択的な情報公開や低リスク分野への研究集中が推奨
- 二重用途リスク
  - プライベートデータの削除やモデルの悪意ある攻撃からの防御を支援する反面、検閲やより強力な攻撃手法開発に悪用されるリスク
  - 用途の管理の必要性
- 解釈可能性技術の過信リスク
  - 解釈可能性技術の能力を過大評価することで、他の重要な安全分野へのリソースが不足したり、AIシステムに対する過信が生じるリスク
  - 厳密な評価やベンチマークを実施し、解釈可能性に関する誤解や誤解釈のリスクを軽減する必要性

Challenges: Research Issues

- 包括的かつ多角的なアプローチの必要性
  - 観察手法と介入手法の調整
  - 特徴レベルの分析と回路レベルの分析による表現とメカニズムの相互作用の解明
  - Intrinsic Interpretability と Post-hoc 分析の組み合わせ

- チェリーピッキングと街灯の解釈可能性（streetlight interpretability）
  - 結果を厳選する傾向（包括的な評価を行わずに少数の説得力のある例や視覚化を議論の根拠にする）
  - 現実的なコンテキストでのみ出現する重要な現象を toy-modelから得られた結果では見落とす可能性がある

Challenges: Technical Limitations

- スケーラビリティの課題と人間依存のリスク
  - モデルのサイズ、タスクの複雑さ、動作の範囲、分析の効率にわたって、実際の AI システムへのメカニズムの解釈可能性のスケーラビリティを実証
  - 現在の解釈可能性の研究では、主観的で一貫性のない人間の評価と、グラウンドトゥルースのベンチマークの欠如が問題になっている
- bottom-up アプローチの課題
  - モデルがより複雑になるにつれて、ニューラルネットワークをボトムアップで完全にリバースエンジニアリングできるかは疑問
- 実環境での分析
- 解釈可能性に対する敵対的圧力
  - モデルを理解するための解釈可能性技術を積極的に不明瞭にしたり誤解を招くような欺瞞的な行動を学習するリスク

Future Directions

Clarifying concepts

- Corroborate or refute core assumptions
- Integrate existing literature and terminology

Setting Standards

- Prioritize robustness over capability advancement
- Establish metrics, benchmarks, and algorithmic testbeds

Scaling Up

- Automation techniques
- Coverage and complexity
- Universality and overarching theories

Expanding Scope

- Vision, multimodal, and RL models
- Top-down and Hybrid
- During and before training

Future Directions Clarifying Concepts / Setting Standards

- 機械論的解釈可能性の基盤となる概念の明確化を目指し、重要な用語や理論的枠組みの整理が必要
  - 既往文献（知見）の統合
  - 中核となる仮説（線型性仮説や重ね合わせ仮説など）の裏付け・反証

- 解釈可能性における評価基準や手法の標準化
  - 能力の向上よりも堅牢性を優先
  - メトリクス、ベンチマーク、アルゴリズムのテストベッドの確立

Future Directions Scaling Up / Expanding Scope

- 技術のスケーリング、特に大規模モデルへの適用可能な手法の開発
  - 複雑なモデルと動作をより幅広く深くカバー

- 範囲の拡大
  - トレーニング中の解釈可能性：
    - トレーニングの前や途中での学習ダイナミクスの解析
  - 多層的分析：
    - トップダウンおよびハイブリッドアプローチの追求
  - 新たなフロンティア（視覚、マルチモーダル、強化学習モデル）：
    - 視覚トランスフォーマーの解析や、強化学習モデルの報酬や目標の表現解析、内部回路変化の研究