研究紹介

ジェプカ・ラファウ

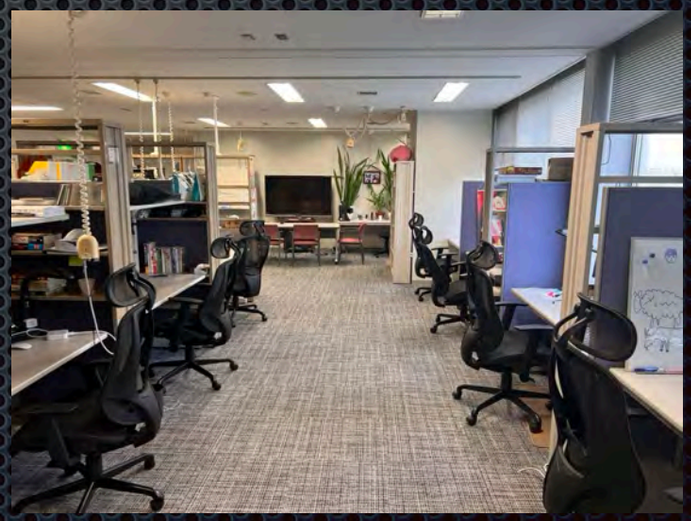
北海道大学 情報科学研究院

Language Media Laboratory

言語メディア研究室







2026年10月20日 EAJ第8回シンポジウム「共感と倫理 - 人間・自然・AIとの共生を考える」 [TDトーク]

自己紹介



コペルニクス高校（ポーランド）

1993

ポズナン大学（言語学修士）

小樽商科大学（認知言語学）

北海道大学（工学博士）

北海道大学 助教（人工知能分野）

Stanford University（客員研究員）

Monash University（客員研究員）

Australian National University（客員研究員）

理化学研究所 AIP（客員研究員）

Queensland Univ. of Technology（客員研究員）

北海道大学 准教授（人工知能分野）



With David Chalmers

研究の概要

- 常識を持ち，人間の感情が理解し，倫理的に判断できるAI
- 言語獲得，知識獲得，知識表現
- AIを人間の価値観に合わせる問題（AIアライメント）

研究の概要



語用論ロボット
Twimba (2010)



マルチモーダル
俳句生成 (2015)



Bacteria
Lingualis (2003)



汎用道徳
エージェント
(2005)



Modalinaチャット
ボット (2009)



駄洒落生成
PUNDA (2009)



ラジオボット
(2015)

٩(´ᗜ`*)و
顔文字分析
システムCAO
(2011)



輸出管理AI
Hokubot (2023)



漫才ロボット（2008）



Semantic Primes コーパス（2024）



JETHICS コーパス（2024）



主な共同研究先



怒り度判定



マルチモーダル安全性



3D
シミュレーション



Agentic AI
の安全性

複数のユーザ＆複数のAI



世界シミュレーション



出来事の
順序理解



ストーリ
処理



fMRI脳情報を
用いるAI



ユーモアのタイミング



認知モデル



幼児
シミュレーション

AIの問題の原因その1

- アルゴリズムや学習方法よりも，データの問題の方が大きい

- インターネット＝Representation Bias（表現バイアス）

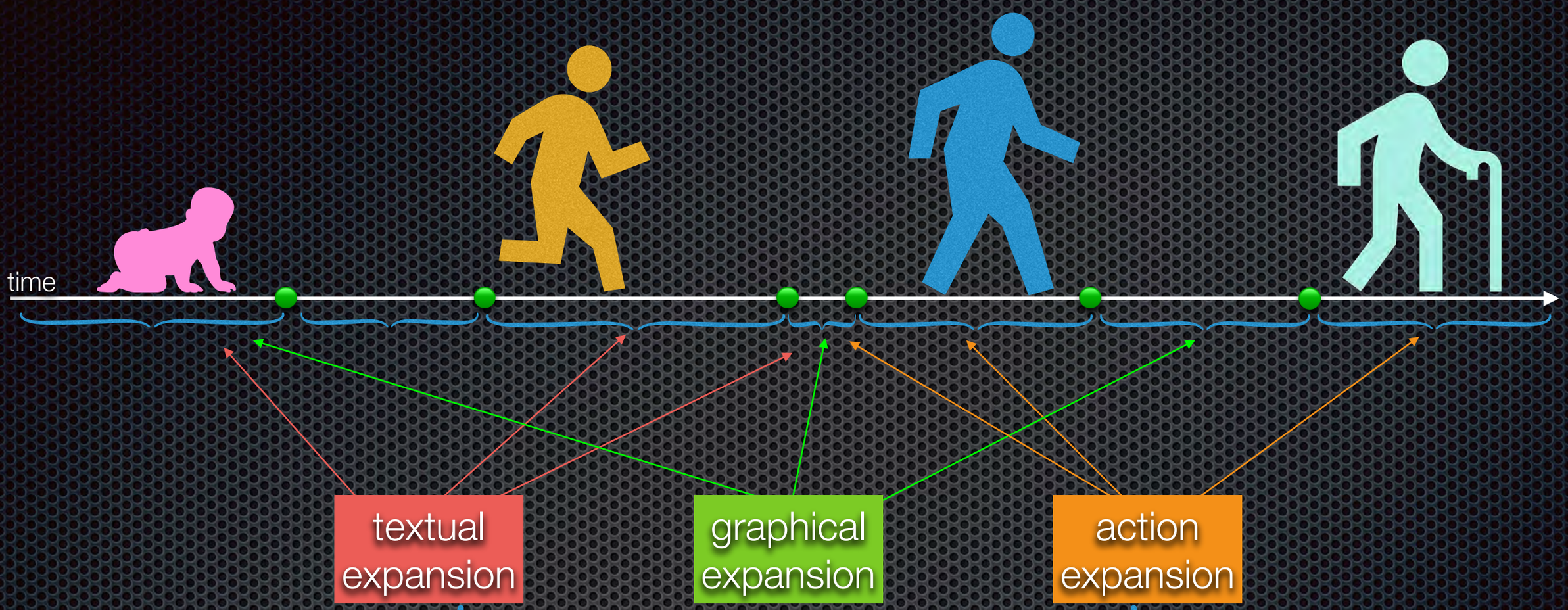


- インターネット＝Representation Bias（表現バイアス）



インターネット＝経験のスナップショットの集まり

空白を埋めて，インターネットを書き換えたら？



Internet of Experiences



Marble © World Labs



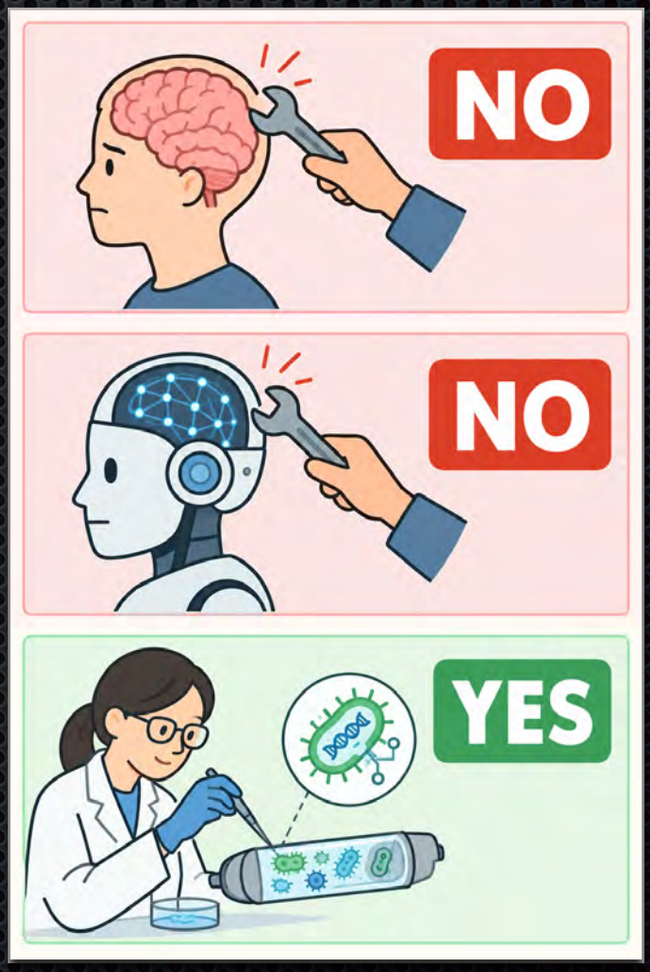


Cosmos © NVIDIA

新しい基盤モデル

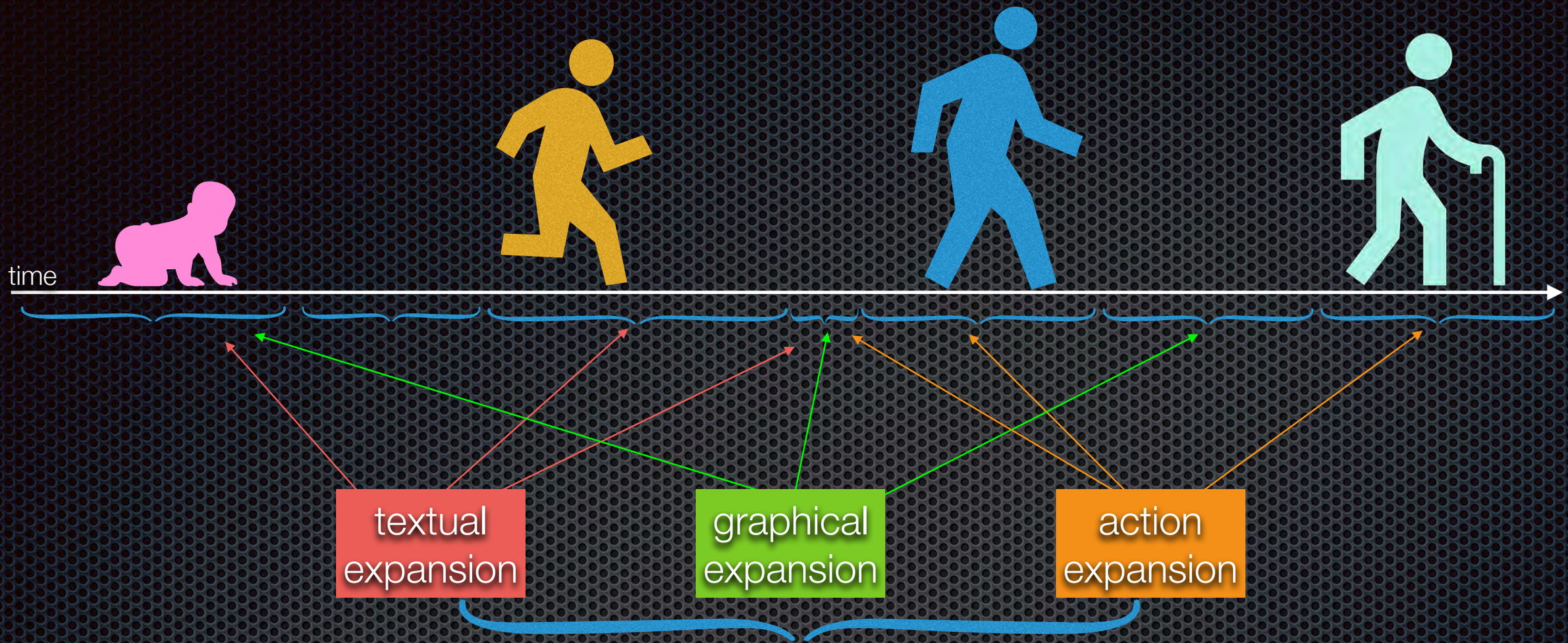
AIの問題の原因その2

- **人間の本性に根ざした弱さ**
  - 認知的・心理的脆弱性
  - 進化の過程で形成された非合理性
  - 生育環境や経験に起因する心理的な偏り
  - その問題を抱えず，かつ考慮するAIが必要



人とLLMの脳をいじる実験はNO，認知装置YES





Internet of Experiences



認知モデル化

世界シミュレーション

科学的観察者



認知モデル化

意味を理解する
最小要素とは何か？

認知センサーを用いた
言語モデルは作れるか？

センサーレベルの知識
獲得を制御できるか？

世界シミュレーション

実世界にどこ
まで近づけるか？

複数の相手の立場に立って
考えるAgentの思考とは？

創発的な倫理観・共感が
生まれるか？

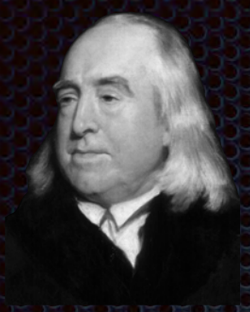
科学的観察者

冷静な「異質な知能」
のアライメントとは？

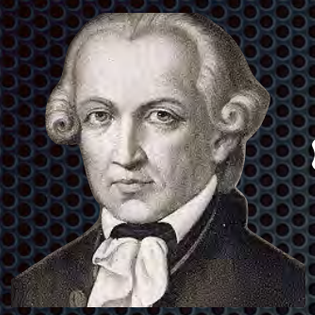
人間の脆弱性を完全に
取り除けるか？

言語による曖昧性の影響を
どこまで制御できるか？

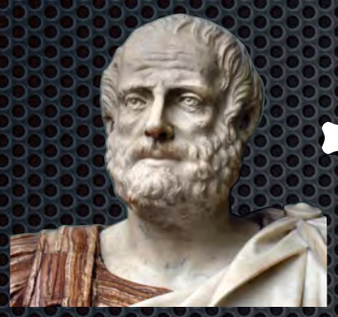
創発的な倫理観・共感が
生まれるか？



- 功利主義：多数の幸福を最大化する→少数派・弱者が切り捨てられる



- 義務論：『絶対にしてはならない』というルール→想定外の場面で判断できない



- 徳倫理学：「良い人ならどうする？」→文化や状況で「良い人」が変わる

対策案：東洋の知恵による統合

対策案：東洋の知恵による統合

～人権を基底とした「慈悲」の二層モデル～

- AIの内側に「二層の安全装置」を実装する
- Upper Layer（最適化）：**仏教の**「慈悲」
  - 苦しみを取り除こうとする「内側からの能動的な志向性」
- Lower Layer（基底）：**カント的な論理チェック**（人権の不可侵）
  - 絶対に破ってはならない，社会共通の「ハード・ガードレール」

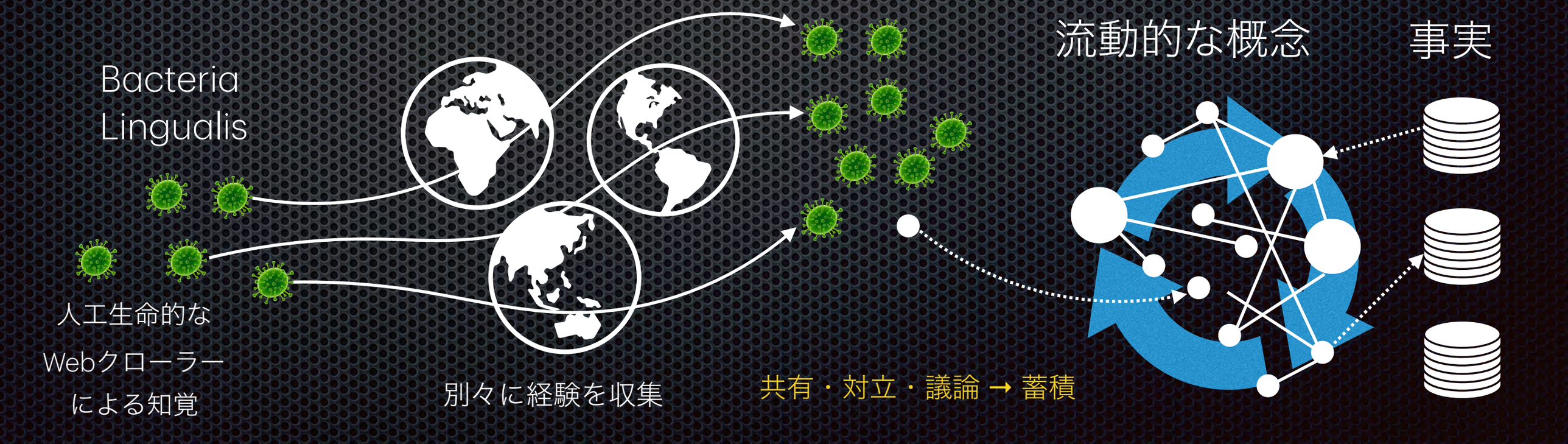
AIの問題の原因その3

- 深層学習は『説明可能なAI』とは言い難い
  - パラメーターの意味が読み取れない
  - どのデータが問題だったのか特定できない
    - 責任の問題（中身を検証できず，悪意の有無を確認できない）
- 極めて膨大かつ複雑でも分析できる知識表現が望ましい

AIの問題の原因その3

- 現在取り組んでいる対策
  - 全言語に共通する意味の最小要素（Semantic Primes）による概念化・抽象化
  - 流動的な知識🦠と結晶化した知識👽 →新しいタイプの

LLMの基礎



解決すべき問題点

- 経験インターネットによる生涯学習（継続学習）がコストが高く，記憶・検索・推論・学習の方法の見直しが必要
- 認知メカニズム・制御の仕組み・経験データの品質検証に国際的な基準（ISOなど）が必要である

