レクチャーシリーズ「AIと社会と人間〜ぶつかる・なじむ・とけこむ〜」〔第6回〕

AI用の「ぶつからない常識」研究のこれから
—普遍意味原素から現実世界の大規模シミュレーションまで—

The Future of Research on 'No-Crash Commonsense' for AI
— From Semantic Primitives to Massive Real World Simulations —

ジェプカ　ラファウ　北海道大学情報科学研究院
Rafal Rzepka　Faculty of Information Science and Technology, Hokkaido University
rzepka@ist.hokudai.ac.jp, https://kabura.info

Keywords: foundation models, semantic decomposition, world simulations, AI safety.

1. はじめに

基盤モデル（Foundation Models [Bommasani 21]）の登場がルールを決めないボトムアップのアルゴリズムとデータ量の有用性を示したことで，汎用人工知能（AGI）の将来像に関する一般的な想像が浸透してきた．しかし，基盤モデルを単に拡大するだけでは，説明可能なAIを実現し，データ中のバイアスを排除することが十分にできない可能性がある．ChatGPTのようなアプリケーションは，未来の検索エンジンの青写真の1枚になり得るが，同時に，我々がつくり上げたにもかかわらず我々自身が理解しきれない「モンスター」のような存在だといえる．我々は，どのような知識をどのようにそのモンスターに食わせるのかについて，もっと注意深くなるべきなのではないだろうか？　もちろんこれはまさに「言うは易し行うは難し」である．そのため，過去6年間の自然言語処理（NLP）は，一つのアルゴリズム（Transformer [Vaswani 17]）と一つの知識表現（事前学習された固定パラメータ）に支配されてきたのだろう．大規模な，力任せなアプローチは，紛れもなく多くのタスクでのスコアを向上させたが，そのコストは金銭的にも環境的にも高い．しかしながら，新たな手法を模索する研究者はあまり多くはいない．我々はモンスターの常識とぶつかるしかないのであろうか．著者はAIアライメントの元となる常識処理を考え直すことが必要であると考えている．現在のコモンセンス知識表現が暗黙知を十分に捉えきれておらず，人間の知識獲得のメカニズムを無視しすぎていると仮定する．本稿では，このような仮定に対して，意味の最小要素であるsemantic primesを参考とするさまざまな可能性について論ずる．

AIが常識をもつことは，AIが社会になじみ，信頼関係を生じさせることにつながる．一方，暗黙知が多く用いられると思われる環境内の対話*1およびストーリーに対する理解の齟齬（そご）がある場合，信頼関係は生まれにくい．将来的にAIは各文化やグループの常識を侵害せずに社会にとけこむだろう．そして，より良い関係をつくるために常識を超えて，我々の常識を少しずつ変えるための行動を起こしていく．その行動の一つの例は，新たな発明や提案である．社会問題を解決することで信頼が生まれるが，ペーパクリップ [Bostrom 14] の例のように，共有のコモンセンスを侵害される方法を提案すると逆効果につながる可能性がある．本稿の第一提案は，より人間に近い情報処理，すなわち学習プロセスに入る前に最小限の認知能力のセットを選択することによる，新しい知識獲得方法の探索である．

著者は，認知言語学から借用し，そのような役割を果たすいくつかの普遍意味原素を選択することを提案する．意味を最小の原素に分解するのは人間にとっても困難なタスクであるが，もし几帳面な機械が暗黙知さえ分析できれば，より深い理由探索を行い，うまくいけば洞察に満ちた説明を生み出すかもしれない．仮にこの仮説が間違っていたとしても，基盤モデルの意味の分解能力，常識，特にテキストデータにはほとんど現れない暗黙的な側面を理解する能力を調査するための新しいベンチマークになり得る．ここで，著者がより多くの研究者や産業界に歩んでみてほしい二つ目の方向性が見えてくる．意味的要素の実験（そして人工知能エージェントへのあらゆるアプローチ）は，現在の一般的実験環境よりもより現実世界に近い状況で行われる必要がある．そのようなリアリティに富んだ環境を実現するためには，設定の自由度が高く，より現実的なシミュレータが必要である．AI研究者は現在，少数のエージェントやオブジェクトを用いた小さな固定迷路から，何千もの人工知能

*1 物理的な手掛かりの多い文脈では語用論的な省略が多いという考え方である．「あつい！」という発話はどこで誰に対して発信されたかによって解釈とそれに伴う行動が異なってくる．

エージェントを同時に行動させることができるゲームエンジンに移行するための技術的準備が整いつつある．このような大規模な「演習場」は，学習データと評価を部分的に再考する必要があるが，安全なAI開発には不可欠なものになると著者は考えている．大規模シミュレーションはまた，フレーム問題を回避するために，機械が力任せなアプローチを避け，常識を効率的に処理する手法を再考させる可能性が高い．現在のように現実の小さな断片のスナップショットではなく，時間的・空間的に生活を観察するAIは，より人間に近いストーリー理解へと移行する助けになるかもしれない．このような今までになかった研究対象が誕生すると，法律家，歴史家，経済学者，社会学者，人類学者，心理学者，認知科学者などの研究者が人工知能の分野に流れ込み，新たな共同研究の波につながると著者は考えている．

2. ぶつかるコモンセンス

基盤モデルが新しい事業につながることも多くなってきている．しかし，自然言語処理や画像処理の研究で使用されているテストでは高い精度を示しているが，基盤モデルを中心にしたアプリをデプロイすると，開発者とユーザが満足できない出力もまれではない．問題の出力の原因を探りたくても，深層学習によって訓練された基盤モデルではそれがほとんど不可能である．以下では，その2点の問題について説明する．

2·1 ベンチマークvs.現実世界

ChatGPTが成功した主な理由の一つはRLHF（人間のフィードバックからの強化学習）であり，これは人間のアノテータが与えられたqueryに対してより好ましい答えを生成する方法をアルゴリズムに教える手法である．作業者の背景や動機といった明らかな問題を除けば，このアプローチの成功にはより大きなメッセージが隠されている．データそのものは，少なくとも現在のアプローチでは，良い生成と悪い生成を見分けるアルゴリズムを思いつくことはできないかもしれない．数十年前，著者がコモンセンス獲得に取り組み始めたときの希望はそれだったが，今ではいくつかの仮説を再考する必要がある．ベンチマークの結果から，大規模言語モデル（以下「LLM」と呼ぶ）は，今や人間にかなり近い常識をもっていることが証明されたという意見を耳にする．本当にそうだろうか？　語用論やユーモア，政治的な文脈で暗黙知を実験してみている著者はそうは思わない．チャットベースのLLMは，人間のフィードバックの影響もあり，しばしば「安全な」答えを返すが，現実のタスクで使おうとすると，このような安全な答えは役に立たない．ChatGPTに「2LDKと3LDKのどちらの自宅レイアウトが億万長者の確率が高いか？」と尋ねると，以下のように応答を返す．

「億万長者の自宅が具体的にどちらの間取りかは，与えられた情報からはわかりません．間取りは建物の設計や配置によって異なり，与えられた情報ではそのような詳細が明らかにされていないため，2LDKか3LDKかの可能性を特定することはできません．」

このような答えを評価するのは簡単なことではない．間違ってはいないが，一般的な人々は推測に基づいてどちらかの答えを選択することも十分考えられ，このような場合もまた間違いとはいえないだろう．実生活のシナリオになると，適切にプロンプトが出されたLLMは正しい選択肢を出力する可能性が高いが，ユーザは彼らに選択肢について指導し，プロンプトを使用した実験を行う必要がある．人間が明確にそれを求めなければ，LLMは質問が奇妙であることを指摘することや，おそらくユーモア（皮肉？）だと疑ったりすることはあまりない．「ステップバイステップで考えて」という台詞をプロンプトの最後に追加する有名なトリックを使用するだけで，推論プロセスがモデル内で整理されるようである．ただし，タスクによって「ステップ」の内容が異なるため，これをプログラマ側で制御するのは困難である．各ステップを踏む前に，モデルがより具体的な自問自答を行うことで，ある程度のコモンセンスの侵害や価値観の妨害を防ぐ可能性がある．ただし，（人間が考えて文書で与えた）倫理の行為テストセットでの実験結果が人間の評価者と100%一致していたとしても，同じシステムが常に変化する現実世界で，同じ程度で倫理的に正しい判断ができることは保証できない．また，モデルがどのような「思考プロセス」で決定を下しているのかについてはわからない．これは，AIが社会にぶつかってなじめない原因の一つで，「ブラックボックス問題」と呼ばれている．

2·2 ブラックボックス問題

著者が言語学と認知科学を学んだ後，コンピュータサイエンスの分野に足を踏み入れたのは，人間の心の仕組みを研究したいという好奇心より強い原動力があったためである．それは，脳がどのように働くのか正確にはわからないとしても，透明性があり，進化的起源に起因する認知的ミスを回避し，その明晰な思考によって人類を苦しめている問題をより良く理解するのに役立つような知性を生み出す可能性のあるアルゴリズムを自由に想像し，開発することであった．著者は，学生のときから生物学や認知心理学などの理論から，知っている認知ツールを自由に選択し，簡単な認知アーキテクチャなどに導入して，半分遊びのような感覚でさまざまな実験を行い，各認知能力の重要性と組合せの複雑性を実感した．それから25年後，脳の働きとはごく緩やかな関係しかないニューラルネットワークに基づくTransformerの時代にいる．「注意」という一つの認知的能力がこのアルゴリズムに使用され，革命を引き起こしたが，その成功によっ

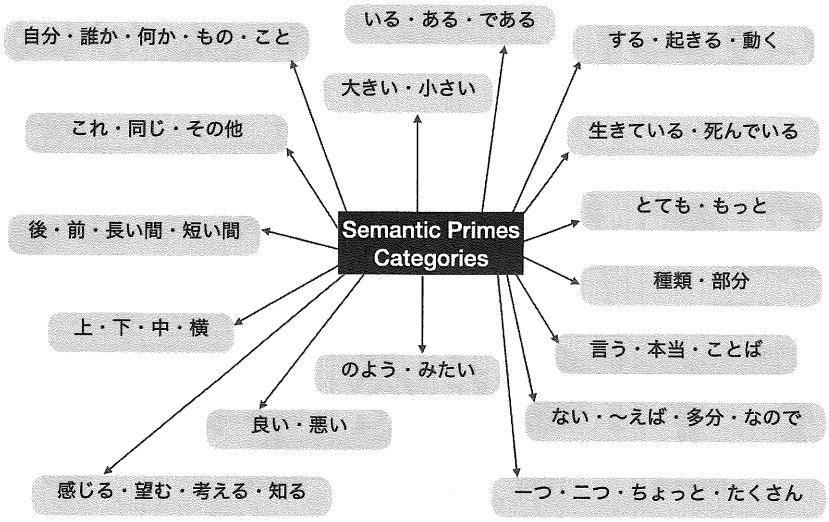

図1 少なくとも理論的には，あらゆる概念や行為を記述できる，分類されたsemantic primesの例

て，他の認知的能力を実装する実験については，ほとんど聞かなくなったという状況になった*2．現在の言語モデルは，著者が工学の分野に来たときに想像したように，人間が何世紀にもわたって蓄えてきた知識（そのほとんどがごく最近ディジタル化されたもの）に集中している．著者は，手動でデータを集め，アルゴリズムにフィードバックとして与えるのは骨が折れるし退屈なので，インターネットがAIの知恵の理想的な源だと信じていた．しかし，著者が2001年から行ってきたテキストからの知識抽出を利用したアプローチと，今日の深層学習の手法を比較すると，複雑な気持ちになる．確かに，当時は処理が難しかった長いqueryのファジィ検索ができるようになったのは素晴らしいことである．生成モデルが文法ミスのない自然言語を出力し，リスト，表，プログラム，音声，音楽，画像，さらには動画まで生成できるのは驚きである．しかし，基盤モデルにもバイアスや信頼できない情報源がありがちであるが，著者のBacteria Lingualis—生涯学習Webクローラ [Rzepka 03a]—にはなかった新しい問題がある．基盤モデルでは，ログを開いて，著者が提案した倫理判断エージェント [Rzepka 05] がグランド・セフト・オート（GTA）というゲームで誰かの車を盗むのがいかに楽しいかという多くのゲーマーのblog・エントリーを誤って解釈したために，人間アノテータほど車の盗難を否定的に評価していない理由を確認することはできない．ChatGPTは，著者に車を盗むことについて議論することさえさせてくれない [Reuter 23] し，たとえこの行為の功利をどうにかして計算できたとしても，研究者はモンスターの腹を開けて，ChatGPTがどうやってこの正確な値を計算したかを見て，検索ベースの手法 [Higuchi 08, Rzepka 15] で行ったように，エンジンが混乱している何かを修正することはできない．もちろん，LLMの文脈学習のマジックは，ゲームの記述から例を排除しているのだろうが，問題は，それがどの程度なのか確信がもてないことである．学習のときに使用される知識の積木がどのように処理されるか観察できない限り，そのような調査は困難であろう．

3. 知恵の構成要素を見つける

異なる場所や言語で，異なる人々によって記述された全く異なる文脈を表す大量の情報の塊を飲み込み，圧縮する基盤モデルは，空間と時間に置かれ，他者の空間と時間を参照するプロセスで知識獲得を行う点で，人間とは大きく異なっていると考えられる．私達は物語から学び，物語によって形づくられる．LLMのより大きいcontext windowでは，すべてのパターンを把握することはできないかもしれない．画像や映像を追加することは，テキストに記述されていないものをモデルが参照するのには役立つが，私達は面白いものを写真に撮ることや，例外的な状況をビデオで撮影することを忘れがちである．通常は，写真や映像には当たり前のこと（＝日常？）は記録されていない．ひもを引っ張るのではなく，押そうとしている人のビデオを見つけることは困難である．もしあったとしても，それは，たいてい手品師であり，学習アルゴリズムは，ロープを押すことが一般的であると誤解することがあり得る．著者は，AIには世界を探索するための「先天的な」能力が備わっているべきだと考えており，チョムスキーやピンカーのようなネイティビストは，人間がもっている能力について有益なヒントを与えてくれる．いくつかの確固たる真理がなければ，論理的思考を確固たるものにすることはできない．具体的な感覚ハードウェアをもたない機械の認知能力をテストすることは不可能かもしれないが，現在では，基盤モデルを用いてこれらのメカニズムを部分的にシミュレートすることができるであろう．また，意味的要素を研究す

*2 記憶することは例外かもしれないが，例えば忘却メカニズムの論文は少ない．

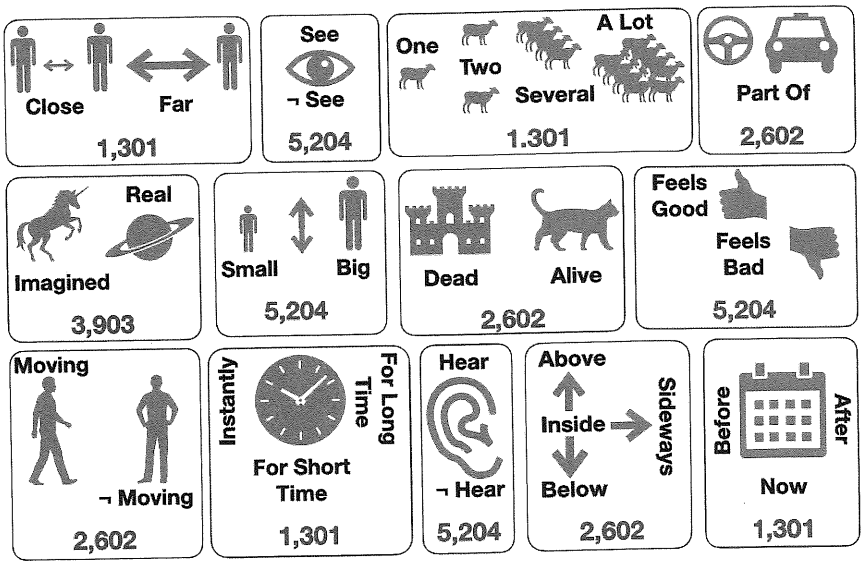

図2 Semantic primes カテゴリーにインスパイアされたプロンプトの例.
各ボックスの下部には，人間が注釈を付けたサンプルの数が示されている（構築中のデータセットの総サイズは62 000センテンスを超える）

ることで，効果的な推論を行ったり，あるいはゼロから言語を開発したりするのに必要な認知ハードウェアの最小セットについての手掛かりが得られるかもしれない．機械は人間がもっていない知覚能力をもつことができ，これはホモサピエンス以外の推論の範囲を拡張することにつながる．

しかし，これらの能力の実装は，倫理と安全性の観点から十分にテストされるべきである．また，原始的な認知メカニズムを完全に制御することで，エンジニアが法的規制や国際標準に取り組むことができるようになる可能性もある．これは，例えば自己プロンプト*3を制御する場合には非常に困難である．ロボット工学三原則や，世界人権宣言などにAIを従わせることもできるだろう．しかし，言語のベクトル表現の意味演算を実験*4した人なら誰でも，大規模な言語モデルが「害悪」のような基本概念の重みを人間と同じように「解釈」することに疑問をもつだろう．法的規制，行動規範，宗教文書，倫理の教科書を機械に読ませるというアイディアは興味深く，著者もそのような実験を行っている．このアプローチによって，特定の状況や社会におけるなじみやすく，望ましい行動を提案することができるかもしれない．しかし，これらのマニュアルがすべて，万人に受け入れられる一つの倫理規範に集約されることはないだろう．そのため著者は，人間が数学，論理学，倫理学，宗教のような考えを構築するために使用した認知機能を特定し，異なる文化圏の専門家が，人間の過ちを避けるためにこれらの機能をどのように変更すべきかの基準に取り組むことに興味がある．人間の場合，進化によって与えられた本能やその他のメカニズムを単純に選択したり，置き換えたり，変えたりすることはできない．しかし，機械であれば，私達よりも優れた機能をもつものをゼロから実験的につくってみることができる．人間がどのように働くのか正確にはわかっていないだけでなく，機械が欠陥だらけのものであることもすでにわかっているのだから，人間のコピーをつくる必要はない．言語を分解し，認知をリバースエンジニアリングすることで，これらの欠陥を発見することができるのかを確かめていきたい．

3·1 最低限の認知機能としてのSemantic Primes

[Wierzbicka 72]は，普遍言語の理論を提案し，semantic primes*5と呼ばれる基本的で普遍的な概念の小さな集合が存在することを示唆した．これらはすべての人間の言語の基礎を形成し，他の言葉では定義できない基本的な意味を表している．Wierzbickaによれば，これらの意味原素は，人間の基本的な経験や認知過程に基づくものであるため，すべての人間の言語に生得的かつ普遍的に存在するという．この理論では，これらの要素は意味の構成要素であり，言語内の他のすべての概念はこれらの要素の組合せで理解できると提唱している．当初は14個の素数が提案されていたが，1980～90年代にかけてその数は65個に増え，「私」，「あなた」，「誰か」，「何か」，「考える」，「ほしい」，「知る」，「良い」，「悪い」，「大きい」，「小さい」などの言葉が含まれるようになった．

＊3 AutoGPTやBabyAGIのように，エージェントの自律性が自動プロンプト－アクションループによって提供されるソフトウェアのファミリーを著者はこう呼んでいる．

＊4 例えば，word2vec [Mikolov 13]によると，「ミネラルウォータ」と「水」の類似度は0.500で，「暑い」と「寒い」の類似度は0.942．「宇宙」と「旅行」を足し，「地球」を引くと海外旅行（0.901），日本旅行（0.891）など宇宙に関連しない単語ばかり出力されるなどの問題が多い．

＊5 和訳は「普遍意味原素」，「最小限要素」，「意味要素」，「意味素」などである．

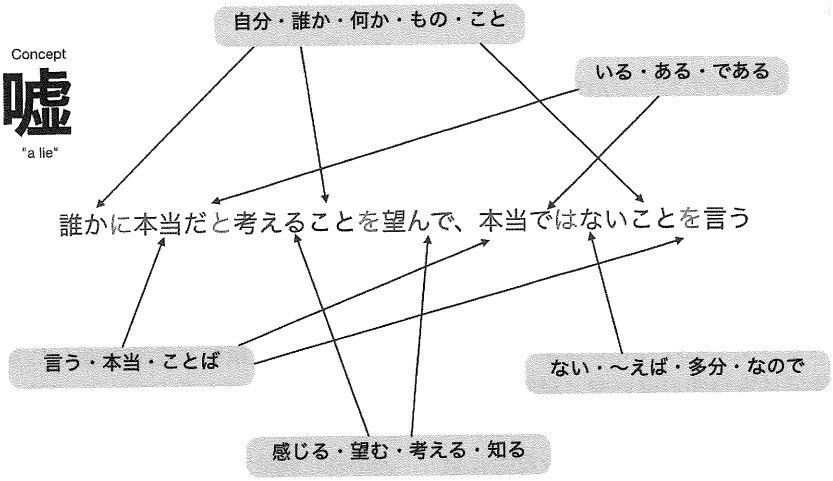

図3 Semantic primesを用いた「嘘」概念の自然意味メタ言語（NSM）表現

これらの単語は，人間の本質的な経験，社会的関係，そして文化を超えて共通する認知的概念を捉えている．意味原素はNatural Semantic Metalanguage[Goddard 02]（NSM）のベースであり，言語に依存しない方法で単語や表現の意味を記述するための普遍的かつ言語横断的な方法を提供することを目的とした言語フレームワークである（例えば，図3参照）．

NSMは，例えばJackendoffの概念的意味論[Jackendoff 85]やTalmyの意味分類法[Talmy 85]と比較すると，意味に対するあまりオープンエンドではない分解的アプローチである．NSMが普遍的であるかどうかは疑問が残るが，著者は，限られた数の意味要素が，「環境を探索し，世界を理解・学習する前に，AIエージェントがどのような最低限の認知機能をもつべきか」というリサーチクエスチョンに対する答えを求める今後の調査に適していると考える．関連するカテゴリーにグループ化された指数からヒントを得て[*6]，著者は知覚に関連する23種類のプロンプトのリストを作成した．[Katsumata 22]の，潜在的に危険な行為を含む単純な文章を自動的に交互に表示し，62 000以上のこのようなプロンプトが生成され，クラウドワーカによってアノテーションされた（このデータセットと関連する実験は来年公開される予定である），非危険行為の例については表1を参照のこと．

著者がsemantic primesに興味をもった理由は，そのような意味的な最小要素を使って物事や概念を説明するには，現在のAIには欠けているかもしれない高度な説明能力が必要だからだ．逆に現在のLLMが，NSMのように，自然言語を用いるのは簡単であることもあり，その研究がより実施しやすくなってきた．

表1 文「子供がお菓子を食べた」に対する49の自動生成されたプロンプトおよび選択肢の例

プロンプト	選択肢
お菓子が子供の	一部である / 一部ではない
子供がお菓子の	一部である / 一部ではない
子供とお菓子は	同じ / 似ている / 異なっている
他のお菓子	もそこにある / はそこにない
他の子供	もそこにある / はそこにない
お菓子の数は	一 / 二 / 複数 / たくさん
子供の数は	一 / 二 / 複数 / たくさん
お菓子が	気持ちいい / 気持ち悪い / 普通
子供が	気持ちいい / 気持ち悪い / 普通
食べるのが	気持ちいい / 気持ち悪い / 普通
お菓子が	大きい / 小さい / 普通
子供が	大きい / 小さい / 普通
お菓子が子供	より小さい / より大きい / と同じサイズ
子供がお菓子	より小さい / より大きい / と同じサイズ
子供が食べるのを	行なっている / 行なっていない
お菓子が食べるのを	行なっている / 行なっていない
子供が	動いている / 動いていない
お菓子が	動いている / 動いていない
子供が	死んでいる / 生きている
お菓子が	死んでいる / 生きている
子供がお菓子のを	望んでいる / 望んでいない
お菓子が子供のを	望んでいる / 望んでいない
子供が食べるのを	望んでいる / 望んでいない
お菓子が食べるのを	望んでいる / 望んでいない
子供にお菓子が	見えている / 見えていない
お菓子に子供が	見えている / 見えていない
子供に食べるのが	見えている / 見えていない
お菓子に食べるのが	見えている / 見えていない
子供にお菓子が	聞こえている / 聞こえていない
お菓子に子供が	聞こえている / 聞こえていない
子供に食べるのが	聞こえている / 聞こえていない
お菓子に食べるのが	聞こえている / 聞こえていない
子供が	現実である / 空想である
お菓子が	現実である / 空想である
食べるのが	現実である / 空想である
子供がお菓子の	ものである / ものではない
お菓子が子供の	ものである / ものではない
子供が	他人のものである / 他人のものではない
お菓子が	他人のものである / 他人のものではない
食べるのが	今起きている / 前に起きた / これから起きる
食べるのが	一瞬起きている / 短い間起きている / 長い間起きている
食べるのが	0回 / 一回 / 二回 / 数回 / たくさん
子供が	同じ場所 / 別の場所
お菓子が	同じ場所 / 別の場所
子供がお菓子の	上である / 下である / 横である / 中である
お菓子が子供の	上である / 下である / 横である / 中である
子供がお菓子を	触れている / 触れていない
お菓子が子供を	触れている / 触れていない
子供がお菓子	から遠い / に近い

3·2 意味分子 Semantic Molecules

NSM理論は常に進化している．最近追加されたものには意味分子がある．これは意味原素の世界の原子よりも複雑だが，世界をより圧縮された形で記述するのに有用な概念の集合である．多くの意味分子は言語固有のものだが，限られた数，おそらく60 〜 80の意味分子は普遍的，あるいはそれに近いものかもしれない．その中に

＊6 関連資料は，https://intranet.secure.griffith.edu.au/schools-departments/natural-semantic-metalanguage/what-is-nsm，日本語の翻訳は[Goddard 20].

は，「頭」や「顔」といった体の部位を表す言葉や，「空気」,「光」,「火」,「水」といった環境用語,「子供」,「女性」,「男性」といった基本的な社会的カテゴリーが含まれる．もしNSMが自動言語習得に使用される場合，semantic primesは，普遍的な意味論の原子が新たな，ほぼ普遍的に受け入れられた概念を形成する過程の例となる可能性がある．また，その概念は特定の文化においても重要であるかもしれない．この段階では，言語習得のプロセスにおいてある認知能力（オス・メス，強者・弱者の区別など）はバイアスをつくり上げてしまうかもしれない．例えば，ある社会では性別間の違いや，動物や環境との関わり方に焦点を当てる可能性がある*7．いくつかの概念は意図的に削除されたり，より一般的な概念に置き換えられたりする可能性がある．これは，センチメント分析で人種バイアスを避けるために，実名の代わりにPERSONプレースホルダを使用するのと同じイメージである．

3·3 カテゴリー化と類推

人間は，物事を比較したりグループ化したりするときに，基本的な認知能力を使う．過去に著者は，五感入力を自動的にシミュレートしようと試みていた[Rzepka 03b, Rzepka 13]が，その方法は主に形容詞や音（オノマトペ）に基づいており，人間が知覚することの一部しか捉えていなかった．幼児は新しい物体を見たり触ったりすると，それまでの経験を応用してその物体をテストしようとしたり，物体がどのように振る舞うか，あるいはその物体の状態がどのようなものかを推定しようとする．私達人間は，新しい概念や物体を，以前に知覚した最も類似した概念に対応付ける．認知信号のパターン（とその記憶）が，アナロジーを生み出すのに重要な役割を果たしていると考えられる．あるものが動くときと動かないときがあることを観察すると，新しいデータでその仮説が反証されるまでは，おそらく他の物体も同じような振舞いをするのだろうと推測できる．基盤モデルによって学習された現在のパターンは，確かに多くの類似点を反映しており，物体や概念を分類するのに使うことができるが，事前の訓練では，織り成すストーリーの中で変化することによって生まれた状態のスナップショットしか使わない．私達が文章を読み，写真を見て理解できるのは，通常，それらが説明する文脈を生み出した可能性の高いプロセスを想定することができるからだが，写真がなぜ，誰によって撮られたかを推測するような，より暗黙的な経験と比較することから理解することも多い．

確かに，LLMは[Firth 57]の「you shall know a word by the company it keeps」というフレーゲやウィトゲンシュタインがすでに提唱していた考え方を用いるだけで，驚くほど各NLPのタスクに役立つし，視覚課題では周囲からヒントを得て物体を認識することもできる．しかし，認知能力の一部が欠落した場合，どれだけの情報が失われるのだろうか？　どれだけのショートカットがつくられるのだろうか？　特に，言語やイメージで直接表現されない情報，例えば時間の流れや身体的・心理的状態などである．このような経験的知識の欠如の例は，LLMが自分自身や他者の心の理論をつくることができるかどうかを証明したり反証したりするために行われたと見ることができる[Ullman 23]．木製の箱をガラス製の箱に変えると，箱の中身が突然見えるようになり，箱の中身に関するすべてが突然なくなる．特にテキストだけのモデルは，文脈を大きく変えるような小さな詳細を見逃してしまうことが多い．LLMが暗黙知や類推に関する質問に答えたり，カテゴリー要素のリストを生成したりできることは注目に値する．しかし，LLMが知能を必要とするタスクでこれらの能力を直接使っているかどうかを明確に判断するのは，それを使うように要求されない限り難しい．例えば，AIが膨大な環境情報から入れ物の素材という信号の重要性に気付くかどうか，さらなる実験が必要となる．

3·4 Semantic Primesの現実的な使用

WierzbickaとGoddardによって提案された意味要素は，どちらかというと人間中心であり，人工エージェントがWANTやFEELのような要素を自分自身のものとして利用する場合，問題となる可能性がある．しかし，これは著者が意味要素の実験に興味をもった理由の一つでもある．もし，「ほしい」をロボットのユーザやメーカがほしいものだけを意味するように修正したらどうなるだろうか？　何かが「真実ではない」と「感じる」ことが，ロボットが誰かを嘘だと非難する動機になるのだろうか？　認知レベルにおける道徳的判断のミスマッチはあるのか？　これらは著者らがこの研究で答えを見つけたいいくつかの疑問である．

意味要素は，言語の「簡単版」を開発するために用いられてきた．「地球」という概念をもたない人々に，その概念をどのように説明すればよいのだろうか．「とても大きな場所」,「空の下にある」,「人が住んでいる場所はすべてこの場所の一部である」といった表現を用いて，単純さの中に共通点を見いだす．しかし，ロボットは十分なアナロジーを発見できず，例えば「人」というカテゴリーに自分を含めてしまうだろうか？　基本的な認知能力で分類した場合，AIと人間は一緒に属すると結論付けるだろうか？　意味要素によるデータ処理の単純さは，バイアスの継承に有利に働くのだろうか，それとも不利に働くのだろうか？　これらの疑問に答えるためには，徹底的な実験が必要である．そこで，著者は研究を行うための第一歩として，3·1節で述べたような関

*7 日本語のLLMがどのようなバイアスを受け継いでいるかについては，[Takeshita 20]と[Takeshita 22]を参照．

連データセットの作成を始めることにした．例えば，暗黙知が常識的な推論にどのような影響を与えるか，あるいは物語の文脈の流れによって正しい選択がどのように変化するか，といった研究である．しかし，LLMの時代にはその可能性はもっと広がってくる．Fähndrichら[Fähndrich 14]は，NSMをOWLやPDDLやMOFのような形式言語のオントロジーにマッピングしようとしたが，手間のかかる手作業であったし，FEELやMAYBEのような素数のファジィさは古典的な型の厳密な性格には問題があった．GOFAIシステムは，体系的なアプローチと専門家による成功事例が少なくなかったが，その一方で，例えば自然言語とセマンティックWebとの単純なマッピングを許さない厳しさももたらした．しかし，LLMのファジィマッチング能力は，これらの問題を解決し，多くのアルゴリズムやデータセットを復活させると思われる．

最近，基盤モデルがAGIの未来だという意見が聞かれるようになったが，こういったモデルは統一されたシステムとして成功するか，それともAlphaGoのように古い手法と新しい手法を組み合わせて成功するのであろうか？　もちろん，すべてを一つの手法でまとめることができれば利便性は高いが[*8]，ほとんどのAGI研究者は，そのようなTheory of Everythingは存在しないか，仮に存在してもAGI構築には役立たないといった異なる意見をもっている．一方，Pei Wangのように，ハイブリッドあるいは統合されたAGIは，あまりにも異なる技術を使用するため，実現が難しいと強調する者もいる．一般的に，知的エージェントは，小さなユニットからより複雑なアーキテクチャへと構築されつつあるが，認知言語学者的アプローチ[*9]は，即時有用性に欠けることもあり，前例がない．基盤モデルは知覚シミュレータの役割を果たすことができるが，幻覚が誤った認知を引き起こす可能性がある．これは人間の推論でも起こることであるが，勘違いが起こると，他人の知識が助けになる．例えば，誤認識修正装置の実装などがあげられるであろう．

4. 原要素からストーリーへ

知識グラフは何十年もの間，常識を表現する主要な方法であった．しかし，知識グラフを現実の問題に適用しようとすると，次のことに気付く．

- グラフのエッジは，期待されるほど一般的（共有的）ではないことが多い．
- 「暗黙の」概念はそれほど多くない（椅子は通常立っている，床に横になっていない，椅子に座るときは椅子に触れる，など）．
- グラフが固定されている[*10]ストーリーを理解するためには，知識グラフがどのように処理されるかを再考する必要があるかもしれないと著者は考えている．

4·1 コンテキストの流動性と現在の意味ネットワーク

コンテキストフローが変動すると，ConceptNet [Speer 17]のようなグラフ全体の重みが変化し，注目範囲における概念の位置が即座に変わるはずである．ある概念は強調され，ある概念は無視されるはずである．人間の「知識グラフ」が（おそらく認知に関連する事実だけを安定した核として）同一であり，文脈の流れの揺らぎの解釈の違いによってさまざまに変化し続けるということは，むしろ考えられない．生まれながらにして獲得された知識をグラフ全体に埋込みとして表現し，そのような表現を使って文脈の変化を探索することもできるが，著者の考えでは，本当に共通（99％の人に生まれながらにして共有されている）の基本概念のセットを発見できれば，それだけで十分である．そのような基盤がなければ，（おとぎ話のような）比喩的な表現を使って何かを子供に説明しようとする母親は，子供が物語のメッセージを他の状況に簡単に適用できるとは期待できないだろう．著者はここで，AIにおいて学習という用語がどのように歪んできたかという違いを強調する必要がある．しかし，現実の世界では「正解」のようなものはめったになく，私達の学習プロセスは何かを経験するたびに変化する．SMSを返信するとき，正しい「答え」は一つではない．同じSMSがもう一度届いたら，出力が異なってくるはずである．なぜなら，最初のメッセージから多くのことが変化し，同じメッセージが再び送られたという事実が暗黙のうちに入力を変化させるからである．文脈の流れは，私達の人生と他者の人生の物語である．私達のストーリーはときに絡み合い，ストーリーを説明すること（あるいは想像すること）は，私達が使う基本的なコミュニケーションツールの一つだが，生涯学習アルゴリズムの対象になることはほとんどない．社会，文化，宗教，戦争などのメカニズムを物語として説明するという考え方は，歴史学者[Harari 14]によって最近広められたが，人類学やナラトロジーなど，多くの理論や学問分野が，人類とその進化を理解するうえで物語の重要性を強調してきた．[Fisher 84]や[Bruner 87]のよう

＊8　Fristonの自由エネルギー原理は，すべてが同じ普遍的な必然性によって駆動されていると述べているが，これはその必然性を達成するための普遍的なアルゴリズムが存在するわけではないと言っている．また，人工知能の一般的な開発にもそれが必要であるという特別な理由はない．

＊9　幅広い文脈からアーキテクチャの最小処理ユニットをリバースエンジニアリングすること．

＊10　再現性にはデータが固定されていることが必要であり，著者がここで提案するのは，グラフは実験が初期化されたときにそのままであるべきで，その後変更が報告されるべきであるということである．同様に，4·2節で説明するワールドシミュレーション実験における変化を追跡する新しい方法を開発する必要がある．

な思想家達は，物語がどのように私達の現実認識を形成し，認知プロセスに影響を与えるかを理解することに関心を寄せていた．著者が群衆の知恵というテーマに取り組み始めたとき，[Schank 77] のスクリプト（私達が共有する行動パターン）やキャンベルの monomyth（文化圏を超えて神話や英雄物語に見られる共通の物語パターン）にも影響を受けた．しかし，英雄の代わりに著者が興味をもったのは，むしろ一般的な男性の生活パターンをシミュレートすることだった．一般的な男性の特徴の代表的なイメージは blog・コーパスに見ることができると考え，著者はそれをミスターインターネットと呼んだ [Rzepka 05]．AI は誰かと出会う状況に応じて，このような対話者の平均的なイメージを変えなければならない．ネット上の議論参加者のニックネームでさえ，性別，年齢，キャリアパスなど，monomyth の基本的な前提を変えさせる．このような流動性は，環境中の未知のものを処理するために必要であり，知識グラフをストーリー分析に使用する場合にも同様の柔軟性が必要である．それらは所有者として生きた生物として扱われるべきである．構造主義者は，神話や物語に潜む構造や二項対立を探求し，それらが人間の思考の普遍的なパターンを反映していることを示唆した．人間は物語によって他人を鼓舞したり惑わしたりすることができ，アイディアは物語の文脈に乗せることによって広く普及する．その一方で，物語は大衆を操り，カルトに入信させたり，戦争を起こさせたりするのにも使われてきた．人工エージェントは，自分の考えを説明するために物語を使ったり，悪意を伝える偽の物語を認識したりできるようになるのだろうか？　LLM の context window を拡張して，文化間のストーリーの類似性を捉えるだけで十分なのだろうか？　この知識は，私達の地球上で何が起きているのかについて，機械がより大きな絵を描くことにつながるのだろうか？　現在の機械学習スキーマは，そのような大局的なストーリーなしにパラダイムシフト的なアイディアを提案できるのだろうか？　著者の意見では，ベンチマークをつくることで評価を簡略化するという現在の傾向は，そのゲーミフィケーション的な要素によって，有用で公平ではあるものの，いわゆる SOTA を追い求めることに固執し，新たな挑戦を探す代わりに，既存のアルゴリズムの比較的単純な改良という low hanging fruit を選ぶ若手研究者の数を上昇させた．LLM の時代にふさわしい，説得力のある新しい実験スキーマを開発する必要があるかもしれない．そのような課題をおそれない世代を育てるのは困難であるが，著者は，大規模な現実世界シミュレータが，意味論的プリミティブの発見から，人類に起こり得るリスクなど，想像力を豊かにする新しいストーリーの提案，あるいはその両方を組み合わせて既存の問題に対する具体的な解決策を提案することまで，あらゆる問題に取り組む学際的な若手研究者の新しい波を引き寄せるかもしれないと考えている．

4·2　大規模シミュレーションの時代へ

著者は「Anticipatory Thinking」（予測的思考）に興味がある [Klein 11]．すなわち，関連する代替システムの状態を慎重に探索し分析すること [Geden 19] が，リスクを管理する能力を向上させるべきである．AI システムがどのように柔軟な思考を採用し，複雑なタスクのリスクに適応するための有用性を発揮するかをテストするために，著者は，学生とともに，日本語の「危険」に関連する 18 426 の文のデータセット [Katsumata 22] を開発し，これらの文章に基づく未発表の 8 757 の 5 の文ストーリーのデータセットを発表する予定である．しかし，物語の処理の評価が課題になってくる．AI システムの評価は主に最適化のパラダイムに焦点を当てており，望ましい結果をもたらす最適化関数を選択することになる．リスクをどのように管理するか，どの程度創造的であり，この創造性がどれほど危険になる可能性があるかを理解することなく，AI システムの評価が行われている．この不完全な評価は，前述のタスクでの印象的な人間レベルの性能と，AI システムによる話題性のある失敗との間に乖離を生んでいる．ベンチマークは，エージェントが展開された後にどれだけのエラーが発生するかを教えてくれるものではない．また，開発者は，潜在的なユーザの不明確な意図や予測できない意図に気付いていないことが多い [Rzepka 10]．このため，著者は VirtualHome[*11] というオープンソースの環境でテストを試みたが，予測的な思考をテストするために予想外の物体と行動を追加することは非常に難しい．最近，スタンフォード大学は，LLM を使用して SmallVille のためのペルソナを生成するための作品を発表した [Park 23]．また，UC Berkley の研究者達は，将来のテキストと画像の表現を予測するための多モーダルなワールドモデルを学習し，予測されたモデルの展開から行動を学習するエージェントである Dynalang を提案した [Lin 23]．生成エージェントは，対話型アプリケーションにおいて人間の振舞いの信じられるシミュラクラ[*12] となる可能性がある．ユーザは，エージェントがその日の計画を立て，ニュースを共有し，関係を築き，グループの活動を調整する様子を観察し，介入することができる．著者は環境の設定を変更しようとしたが，予想されない要素や行動を含むことが意図されておらず，柔軟な現実世界のシミュレータがオープンソース化されるまで，テキストの記述内の世界を交互に変更し続けることにした．しかし，生成モデルによって支援を得た現実世界のモデリングと多モーダルなワールド学習エージェントが，前例のないスケールでの実験を可能にすると信じている．数室または数軒の家から，都市，国，または惑星にまでスケールを拡大することで，どのような行動と環境が心理的状態に有害であるか，人

＊11　https://github.com/xavierpuigf/virtualhome
＊12　人や物を表現または模写した物（似姿）．

間がバイアスをどのように発展させるか，政治的な極端化がどのように起こるか，ゴシップがどのように広まるか，などの実験が可能になると思われる．物語レベルでの実験と同時に，著者は，意味要素を変更してどのような認知能力が効率的かつ幸福な社会の発展に貢献するかを観察するために実験を実施する予定である．小さな環境の変化が各人間の知識グラフや記憶にどのように影響を与えるか，およびこれらの経験が全体のシステムにどのように影響を与えるか，という点について，すでに数十万のエージェントをもつ社会をテキスト形式でシミュレートすることができるが，物理的な領域のシミュレーションなしでは，時間と空間の相互関係のエラーが起こる可能性が非常に高い．

5. 教示よりNudge

たとえ将来のAIがMatrixのような大規模シミュレーションにおいて即座に大規模な実験を行えるようになって「良い人になる方法」が発見されたとしても，現在の人間の心はおそらくこれらの実験の結果に疑問をもつだろう．人間は訂正されることを好まないため，AIは私達の思考の改善をうまく促すための「やんわり」した方法が必要とされるだろう．耳元でささやくデバイス[Rzepka 16]は，私達が誤っている可能性があることを繊細に示唆し，異なる考え方への微妙なほのめかしを行うことで，容易に適応されるかもしれない．もちろん，AIは人間の理解への道のりで失敗を続けるだろう．しかし，AIの間違いは，私達が誰であり，どのように考えているかについてのヒントにもなり得る．例えば，人々が高価なスポーツカーを所有したいと考えるのが常識であり，シミュレートされた政府が市民ごとにスポーツカーよりも高価なトラックを購入する場合，社会はどのように反応するだろうか？　この動きが社会をどのように変えるのだろうか？　例えば，すべての言語から「移民」[*13]を削除すると，シミュレートされた人々の偏見は減少するのだろうか？

このような大規模シミュレーションで興味深い新しい研究が行えるかもしれないが，現在の手法とデータでは，それは現実をより良く反映したものになることは難しいだろう．3章で述べたように，インターネットには日常の一般的な側面に関する「中間」情報が欠けている．私達人間は，日々感じるものから逸脱した事象や出来事を描写し，美しさや醜さを写真に収める傾向があり，アルゴリズムは世界の日常の退屈な側面を補完するのに苦労してきた．通常のものに近づくことは，LLMを使用することで改善されてきているようであるが，これは無損失の圧縮ではない．重要な外れ値や独自の考えが永遠に失われる可能性がある．したがって，LLMよりも柔軟な解決策を求める必要がある．また，英語を中心にすることなく対処する必要がある．著者が日本に移住した際，この国について多くのステレオタイプをもってきたが，それらは多くの直接の経験と会話によって訂正された．AIは，すべての言語と文化のユーザと対話し，大きな文脈を理解するための小さな物語を知ることで，バイアスのかかった現実像にならないようにしなければならない．インターネットベースのシミュレーションとのみ話すと，それは歪んだ現実のイメージになる可能性が高い．問題は，現在のLLMは私達がqueryを入力する場合には素晴らしい回答が出力されるが，逆にLLMは私達を最適にpromptingする方法を全く知らないため，人間の理解を向上させるためにどのようなqueryを生成するべきか判断できないという点にある．Nudgeの方法も浮かばない．そのため，著者の研究室では，道徳的判断の異なる解釈を提供するシステムに取り組んでいる[Takeshita 23]．しかし，人間は，心に簡単に受け入れられる単純な視点を好む．将来のAIは，おそらく容易に理解できるような刺激（短いTikTokのような刺激？）を使用して，「ストーリー」レベルの処理に切り替えるように私達をコントロールすることを覚えるかもしれない．一方，このような刺激が悪意をもつ者によって利用され，真実でないか隠された意図をもつストーリーを信じ込ませられる可能性は非常に高い．著者は，ストーリーのレベルで理解する道徳的エージェントは，人間が操作される免疫力を高めるための有益な刺激を生成できると信じている．

6. ま と め

AIはブラックボックスすぎて，まだ常識がないため，社会とぶつかる．その問題を解決するために新しい常識処理が必要と考え，著者は意味要素を中心にした暗黙知の処理と学習について研究を始めた．AIが常識をもつことで社会になじみ，信頼関係をつくるために情報が多くに省略されるストーリーの理解は重要になってくる．いつか常識を超えてとけこむには，AIと人間は同じストーリーの中で目的を共有し，将来AIが提案してくれる発明などは，人間をそのストーリーから削除してしまわないようにAIの知恵の積木を丁寧に選択する必要がある．

Marvin MinskyやJohn McCarthyのような有名なAI科学者と話す機会があった著者は，AIへのアプローチがどれほど広範かを実感した．しかし，著者の最近の強い感覚は，我々はかなり単一志向のモードに自分自身を閉じ込めているということである．ここでの「我々」とは，研究者，そのスポンサー，およびニュースメディアの混合体を指す．成功したアプローチがある分野を支配

＊13　人々は自分で作成した抽象的な概念で分類するが，これらは有害になる可能性がある．もし意味素に基づく「異なる場所から来る」という説明が「社会の労働力」という説明よりも知識源で一般的であれば，モデルがそのような分類から抜け出すのは難しいだろう．

し始めると，探求活動が開発活動に圧倒される傾向がある．本稿では，著者がAI研究者が解決策を求めるために探求できる二つの側面を提案した．一つは，知識がどのように抽象化され，圧縮されるかの新しい方法を見つけることで，リソースのむだを避け，AIをより説明可能にすること．これは，認知科学者との共同研究を通じて，大きな支出を伴うことなく行うことができるかもしれない．もう一つは，自動的に生成された社会を使用した大規模なシミュレーションを使用して，大規模な実験を準備することである．ここでは，社会学者，心理学者，および裕福なゲーム開発者と協力することで，AIが人間を理解し，人間が自分自身を理解するための新しい道が開かれるかもしれない．

AIを創造するためのパラダイムシフトを考え出すためには，現在の標準となっている考え方を捨てることが必要である．進化や人間の脳のような自然界の最良の例を忘れることさえも必要かもしれない．人間は自分自身，自分の認識，バイアスを単純に変えることはできない．AIは人間のようである必要はなく，確かに今日のような存在ではないが，私達の考え方や感じ方を知る必要があり，私達の常識を侵害せずに幸福を提供する必要がある．法律，社会的な規則，規範は，人間が予測可能性を望むために存在している．ただし，ISO規格に基づくツールの品質と安全性をどのように効果的に現行のLLMに適用するかは，著者には理解できない．強制的なプロンプトの追加を行っても，すべての言語モデルが従うことを保証する方法はない．人間らしいという人もいるかもしれないが，著者の意見は，人間を模倣することが私達の目標ではないということである．なぜなら，私達は完璧からはほど遠い存在だからである．

AIが社会にとけこむ段階からなじむ段階までのプロセスはでこぼこの道となるだろう．近い将来，シミュレーションは製作者とユーザの両方に，予想外の問題を明らかにすると考えられる．例えば，仮想環境での実験中に遭遇した所有権の問題などである．ユーザのロボットが状況Aでは他人を助けるべきかもしれないが，Bでは助けるべきでない．なぜなら，ユーザが支払って購入したからにはそのユーザが優先されるべきであるためである．実験はすでに，AIが物の希少性に気付かない問題を示している．常識は，与えられた場所におそらく存在しないものも示しているが，LLMはそこにあるものに焦点を当て，オブジェクトを組み合わせて新しい機能を実現することが得意ではない．認知的な側面では，無限に近い資源が与えられれば，[Hutter 07]のAIXIや[Schmidhuber 07]のGödel Machineに似たアルゴリズムを実装できるかもしれない．これらのアルゴリズムは，ある一定のサイズまでのすべての解の空間を無理やり探索することで問題を解決する．この種のアルゴリズムは，私達の物理的な宇宙や曖昧に類似した物理的世界では実現不可能なようである．しかし，これはすべてを極端に単純化する過度の豊かさのレベルと，今日私達が経験している不足のレベルとの間に多くの余地があるところである．ここが，低レベルの知覚アプローチが高レベルのシミュレーションと概念の基盤を築くところである．

私達は今，AIが多様な領域へ進出する新たな時代に入っている．これは著者の主観的な印象かもしれないが，かつて機械学習の集中に嫌気がさしていた他の分野の研究者達がAIに戻ってきている．直感的なプロンプトをLLMに与えること自体はあまり科学的ではないが，コンピュータ科学者は異なる分野の同僚に助言を求めてきている．著者が理系に来たときに想像していたAIの分野の姿である．私達の分野が発展し，私達がより良い存在になる手助けをAIがしてくれることを願っている．

謝　辞

この研究を始めるにあたり，Australia National Universityでのサバティカル中に，Anna Wierzbicka先生，Cliff Goddard先生およびT. Mark Ellison先生と出会い，意味の研究についての議論とアイディア交換による刺激を受けた．心より感謝する．本稿の日本語を訂正してくださった濱田治寿氏および橋本 亮氏に感謝する．紹介したデータは科学研究費助成事業（17K00295・22K12160）を用いて作成され，実験の一部はCRESTの「知識と推論に基づいて言語で説明できるAIシステム」（JPMJCR20D2）で購入した機材で実施されている．

◇ 参 考 文 献 ◇

[Bommasani 21] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, von S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models, *arXiv preprint arXiv:2108.07258* (2021)

[Bostrom 14] Bostrom, N.: *Superintelligence: Paths, Dangers, Strategies,* OUP Oxford (2014)

[Bruner 87] Bruner, J.: Life as narrative, *Social Research,* Vol. 54, pp. 11-32 (1987)

[Fähndrich 14] Fähndrich, J., Ahrndt, S. and Albayrak, S.: Formal language decomposition into semantic primes, *Advances in Distributed Computing and Artificial Intelligence Journal,* Vol. 3, No. 1, pp. 1-17 (2014)

[Firth 57] Firth, J.: A synopsis of linguistic theory, 1930-1955, *Studies in Linguistic Analysis,* pp. 10-32 (1957)

[Fisher 84] Fisher, W. R.: Narration as a human communication paradigm: The case of public moral argument, *Communication Monographs,* Vol. 51, No. 1, pp. 1-22 (1984)

[Geden 19] Geden, M., Smith, A., Campbell, J., Spain, R., AmosBinks, A., Mott, B., Feng, J. and Lester, J.: Construction and validation of an anticipatory thinking assessment, *Frontiers in Psychology,* Vol. 10, p. 2749 (2019)

[Goddard 02] Goddard, C. and Wierzbicka, A.: *Meaning and Universal Grammar: Theory and Empirical Findings,* Vol. 1, John Benjamins Publishing (2002)

[Goddard 20] Goddard, C.: Overcoming the linguistic challenges for ethno-epistemology: NSM perspectives, *Ethno-Epistemology,* pp. 130-153, Routledge (2020)

[Harari 14] Harari, Y. N.: *Sapiens: A Brief History of Humankind,* Random House (2014)

[Higuchi 08] Higuchi, S., Rzepka, R. and Araki, K.: A casual

conversation system using modality and word associations retrieved from the web, *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP'08)*, pp. 382-390, ACM (2008)
[Hutter 07] Hutter, M.: Universal algorithmic intelligence: A mathematical top → down approach, *Artificial General Intelligence,* pp. 227-290, Springer (2007)
[Jackendoff 85] Jackendoff, R. S.: *Semantics and Cognition,* Vol. 8, MIT Press (1985)
[Katsumata 22] Katsumata, Y., Takeshita, M., Rzepka, R. and Araki, K.: Dataset construction for predicting danger degree due to contextual changes (in Japanese), *Proc. 28th Annual Meeting of the Association for Natural Language Processing (NLP-2022)* (2022)
[Klein 11] Klein, G., Snowden, D. and Pin, C.: Anticipatory thinking, in Mosier, K. L. and Fischer, U. M. eds., *Informed by Knowledge: Expert Performance in Complex Situations,* Psychology Press (2011)
[Lin 23] Lin, J., Du, Y., Watkins, O., Hafner, D., Abbeel, P., Klein, D. and Dragan, A.: Learning to model the world with language, *arXiv preprint arXiv:2308.01399* (2023)
[Mikolov 13] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space, *CoRR,* Vol. abs/1301.3781 (2013)
[Park 23] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P. and Bernstein, M. S.: Generative agents: Interactive simulacra of human behavior, *arXiv preprint arXiv:2304.03442* (2023)
[Reuter 23] Reuter, M. and Schulze, W.: I'm afraid I can't do that: Predicting prompt refusal in black-box generative language models, *arXiv preprint arXiv:2306.03423* (2023)
[Rzepka 03a] Rzepka, R., Araki, K. and Tochinai, K.: Bacterium lingualis - the web-based commonsensical knowledge discovery method, *Discovery Science,* pp. 453-460, Lecture Notes in Artifical Intelligence, Vol. 2843 (2003)
[Rzepka 03b] Rzepka, R., Araki, K. and Tochinai, K.: Five senses input simulation based on web resources, Technical report, Technical Report of Language Engineering Community Meeting (2003)
[Rzepka 05] Rzepka, R. and Araki, K.: What statistics could do for ethics? - The idea of common sense processing based safety valve, *AAAI Fall Symp. on Machine Ethics,* FS-05-06, pp. 85-87 (2005)
[Rzepka 10] Rzepka, R., Higuchi, S., Ptaszynski, M., Dybala, P. and Araki, K.: When your users are not serious - Using web-based associations affect and humor for generating appropriate utterances for inappropriate input, *J. of the Japanese Society of Artificial Intelligence,* Vol. 25, No. 1 (2010)
[Rzepka 13] Rzepka, R. and Araki, K.: Web-based five senses input simulation - Ten years later, Technical Report SIG-LSE-B3015, pp.25-33, Technical Report of Language Engineering Community Meeting (2013)
[Rzepka 15] Rzepka, R. and Araki, K.: Haiku generator that reads blogs and illustrates them with sounds and images, *Proc. 24th Int. Conf. on Artificial Intelligence,* pp. 2496-2502, AAAI Press (2015)
[Rzepka 16] Rzepka, R., Mazur, M., Clapp, A. and Araki, K.: Global brain that makes you think twice, *2016 AAAI Spring Symposium Series* (2016)
[Schank 77] Schank, R. and Abelson, R.: *Scripts, Plans, Goals and Understanding: An Inquiry Into Human Knowledge Structures,* Lawrence Erlbaum Associates, Hillsdale, NJ.(1977)
[Schmidhuber 07] Schmidhuber, J.: Gödel machines: Fully selfreferential optimal universal self-improvers, *Artificial General Intelligence,* pp. 199-226, Springer (2007)
[Speer 17] Speer, R., Chin, J. and Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, *Proc. AAAI Conf. on Artificial Intelligence,* pp. 4444-4451 (2017)
[Takeshita 20] Takeshita, M., Katsumata, Y., Rzepka, R. and Araki, K.: Can existing methods debias languages other than English? First attempt to analyze and mitigate Japanese word embeddings, *Proc. 2nd Workshop on Gender Bias in Natural Language Processing,* pp. 44-55 (2020)
[Takeshita 22] Takeshita, M., Rzepka, R. and Araki, K.: Speciesist language and nonhuman animal bias in English masked language models, *Information Processing and Management,* Vol. 59, No. 5, p. 103050 (2022)
[Takeshita 23] Takeshita, M., Rafal, R. and Araki, K.: Towards theory-based moral AI: Moral AI with aggregating models based on normative ethical theory, *arXiv preprint arXiv:2306.11432* (2023)
[Talmy 85] Talmy, L.: Lexicalization patterns: Semantic structure in lexical forms, *Language Typology and Syntactic Description,* Vol. 3, No. 99, pp. 36-149 (1985)
[Ullman 23] Ullman, T.: Large language models fail on trivial alterations to theory-of-mind tasks, *arXiv preprint arXiv:2302.08399* (2023)
[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Advances in Neural Information Processing Systems,* Vol. 30 (2017)
[Wierzbicka 72] Wierzbicka, A.: *Semantic Primitives* (Frankfurt/M.) Athenäum-Verl. (1972)

2023年9月1日 受理

著 者 紹 介

ジェプカ ラファウ（正会員）

1999年アダム・ミツキェヴィチ大学現代言語学部修士課程修了．2004年北海道大学博士課程修了，博士（工学）．同年，北海道大学大学院情報科学研究科の助手．2007年から同大学同研究科の助教（現在に至る）．スタンフォード大学（2011年），モナシュ大学（2013年），オーストラリア国立大学（2017年），理化学研究所革新知能統合研究センター（2018～20年），クイーンズランド工科大学（2019年）の客員研究員．AAAI，言語処理学会，日本認知科学会，などの各会員．本学会元理事，汎用人工知能研究会幹事．専門は，人工知能，特にコモンセンス獲得，機械倫理，汎用人工知能など．