AI 哲学者: 哲学論文の自動生成に向かって

竹下昌志 [1]　松田新 [2]
[1] 名古屋大学大学院情報学研究科　[2] 北海道大学大学院文学院
takeshita.masashi.68@gmail.com　matsuda.arata.philosophy@gmail.com

概要

大規模言語モデル（LLM）を用いた研究自動化は急速に進展しているが、その対象は主として機械学習や数学などに限られており、哲学を含む人文学分野での検討は依然として不足している。本研究では、哲学論文を自動で生成する **AI 哲学者**を提案する。本手法は、特定の哲学論文を入力として、分析、アイデア生成、論文執筆までの一連の研究プロセスを自動的に実行し、その論文に対するコメンタリ論文を出力する。評価実験では、LLM による自動評価を通じて、生成されたコメンタリ論文の品質を検証した。その結果、AI 哲学者は一定水準以上の論証の一貫性と構成を備えた論文を生成できることが示された。

1　はじめに

大規模言語モデル（LLM）の発展により、研究活動の複数の工程を LLM によって自動化することが現実的になりつつある [1]。特に AI 研究においては、アイデア生成から実験、論文執筆までを End-to-End で実行する手法が提案され、国際会議ワークショップの査読を通過する水準に達している [2]。

しかし、既存研究の多くは、アイデア生成や実験、論文執筆といった個別の研究ステップに焦点を当てており、研究プロセス全体を全自動化する試みは依然として限られている [1]。また，これらの自動化研究は主に「AI for Science」の文脈で進められており、対象分野の多くは機械学習を含む経験科学に集中している。一部には他分野への展開も見られるものの、とりわけ人文学における研究自動化の検討は極めて限定的である。

本研究では、こうした状況を踏まえ、哲学分野に焦点を当てた研究自動化手法として **AI 哲学者**を提案する。AI 哲学者は、対象となる哲学論文を入力として受け取り、その論文に対するコメンタリ論文を、全工程を通じて自動生成する手法である。本手法は、これまで主として経験科学を対象としてきた研究自動化の枠組みを人文学分野へと拡張する試みである。さらに本研究では、AI 哲学者によって生成された論文を LLM により評価する実験を行い、提案手法の有効性を検証する。

本研究の主な貢献は以下のとおりである。

- コメンタリ論文生成に必要なほぼすべての工程を LLM によって実行する、哲学研究の全自動化手法として AI 哲学者を提案する（3 節）。
- 提案手法によって生成された哲学論文に対する評価実験を実施し、その有効性を定量的に検証する（4, 5 節）。

2　背景：哲学研究の自動化

本研究で AI 哲学者を提案するにあたり、哲学研究の自動化を動機づける理由を述べる。

2.1　哲学研究と科学研究の相違点

既存研究（7 節参照）では科学研究の自動化に焦点があたっており、哲学研究の自動化は限定的であった。その理由として、哲学という分野の以下のような特徴が考えられる。

- 先行研究への自動アクセスが困難である。背景には、オープンアクセス化の遅れ、アブストラクトや本文が標準化されておらず（いわゆる IMRAD 形式[1)]ではなく）論文の内容が推定しづらいことなどがある。
- （先行研究への自動アクセスが困難であるにも関わらず）先行研究への詳細な参照が求められる。単に参考文献として言及するだけでなく、しばしば直接引用を含む検討が求められる。

こうした事情から、哲学においては、先行研究を適切に参照しながら論文を執筆するような自動化手法の構築は困難である。本研究が特定の論文のみに注

1）　科学論文の標準的な構造である **I**ntroduction, **M**aterials and methods, **R**esults, **A**nd **D**iscussion の頭文字をとったもの。

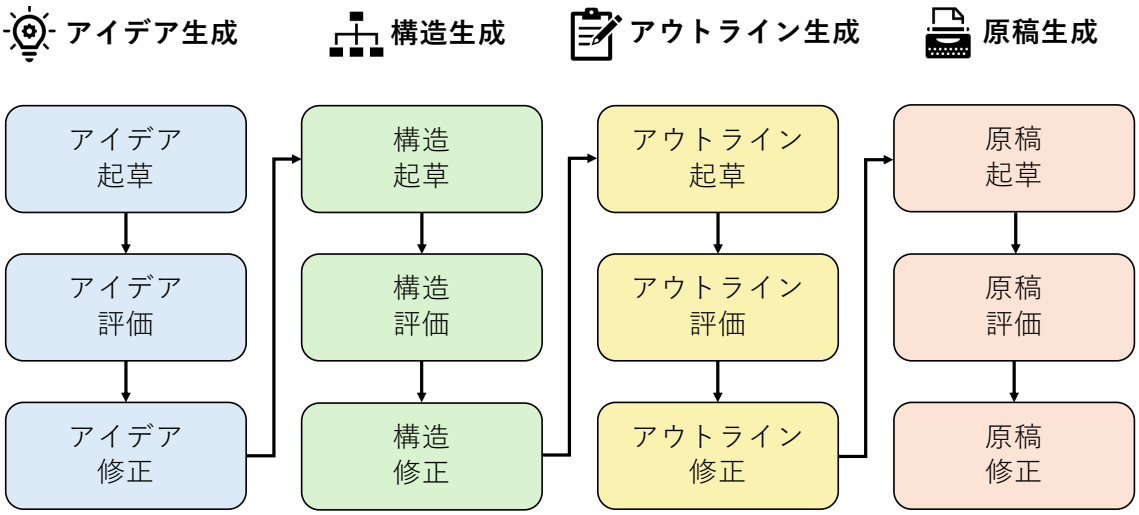

図 1　AI 哲学者の全体処理フロー。四段階（アイデア生成・構造生成・アウトライン生成・原稿生成）からなり、各段階は三つのステップ（起草・評価・修正）で構成される。

目して要約と批判的検討を行うコメンタリ論文の生成を目指すのもこのためである。

一方で、哲学研究の自動化が行いやすい側面もある。多くの場合、実験や観察が不要であり、論文の大半が理論的な考察である。さらに、図表が用いられることも少なく、論文の大半が自然言語での議論である。これらは、物理的実体を持たないために実験することが困難だが自然言語の扱いに優れた LLM による自動化において有利な特徴だと考える。

2.2　哲学研究の自動化による利益

哲学研究の自動化には、二つの潜在的利益がある。第一に、哲学研究のリバースエンジニアリングを可能にする。しばしば哲学研究は、経験豊富な研究者による「職人技」のようなものとして捉えられ、研究プロセスが分析されることは少なかった [3]。しかし自動化のために研究プロセスを分解・実装することで、研究プロセスが分析可能なものとなり、さらなる発展・効率化への道が開ける。

第二に、新たな価値観・世界観を AI が提案する可能性を示唆する。哲学はその歴史を通して、様々な価値観・世界観を世界を探索してきたが、この探索は人間の研究者の時間や注意などの制約のもとにあった。しかし哲学研究の自動化により、これまで十分に検討されてこなかった価値観・世界観が AI から提案される可能性が開ける。

3　AI 哲学者

AI 哲学者は、対象となる哲学論文を入力として受け取り、その論文に対するコメンタリ論文を自動生成する手法である（図 1）。

AI 哲学者は、(1) アイデア生成、(2) 構造生成、(3) アウトライン生成、(4) 原稿生成の四段階から構成される。各段階は、起草、評価、修正の三つのステップからなり、起草で生成された出力を評価し、その結果に基づいて修正を行うという反復的な改善過程をとる。この設計は、The AI Scientist [4, 2] に代表される自動研究手法の枠組みを参考にした。使用したプロンプトの概要を付録 A に示す。

本研究では、各段階につき起草・評価・修正を 1 サイクル実行し、修正後の再評価は行わない。複数回の反復による品質向上は今後の課題とする。

3.1　アイデア生成

アイデア生成段階では、対象論文に対する五つの異なるコメンタリーアイデアを生成する。各アイデアは以下の四つの構成要素を持つ：(1) 焦点：対象論文のどの部分・主張・前提に着目するか、(2) 主要論題：コメンタリーの中心的主張、(3) 論証：主張を支える議論の概要、(4) 意義：より広い哲学的議論における貢献。アイデア生成には四種類の戦略のいずれかを用いる：推論批判（対象論文の推論の誤り等を指摘）、前提・原理批判（中心的な前提や原理を批判）、含意の拡張（さらなる哲学的含意を議論）、異論と応答（対象論文で検討されていない異

論の提示とそれへの応答)。評価ステップでは、五つのアイデアそれぞれについて関連性、独創性、議論の強さ、意義の四観点から評価する。修正ステップでは、最も評価が高かったアイデアを精緻化し、明確化や想定反論への応答を追加する。

3.2 構造生成

構造生成段階では、アイデア生成段階で最も高く評価されたアイデアに基づき、コメンタリ論文全体の段落レベルの構造を設計する。生成される構造は、(1) 序論 (4–5 段落)、(2) 本論 1：対象論文の説明 (9–12 段落)、(3) 本論 2：主張の展開 (17–20 段落)、(4) 結論 (2–3 段落) の四セクションから成る。起草ステップでは、各段落について要約と論文全体における役割を生成する。評価ステップでは、論理的飛躍、冗長性、矛盾といった問題点を検出し、修正ステップにおいてこれらの問題を改善した構造を生成する。

3.3 アウトライン生成

アウトライン生成段階では、構造生成段階で設計された各段落について、主題文と展開メモを作成する。主題文は段落冒頭に置かれる一文であり、当該段落の中心的メッセージを示す。展開メモは、主題文に続く議論の展開方法を指示するガイドである。

本段階では、構造段階で生成された段落構造に厳密に従うよう LLM に指示し、セクションや段落の追加、削除、並べ替えを行わない。評価・修正ステップでは、論理的飛躍や不適切な言い換えを検出し、必要な修正を加える。

3.4 原稿生成

原稿生成段階では、アウトラインに基づいて完全なコメンタリ論文原稿を生成する。生成する原稿は、コメンタリ論文の標準的な単語数に合わせおおよそ 3,000〜4,000 語の原稿を生成するよう LLM に指示する。また、各段落は指定された主題文で開始し、展開メモに従って議論を展開する。

本段階でもアウトラインへの厳密な遵守を LLM に指示し、主題文の言い換えや内容の追加、削除、再編成を行わない。評価ステップでは、論理的飛躍や用語の不適切な言い換えを検出し、修正ステップで最終原稿を完成させる。

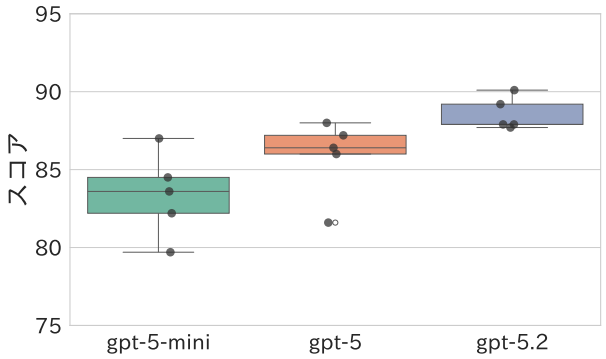

図 2 gpt-5.2 を査読者モデルとした場合の各モデルの生成原稿の評価スコアの結果 (各モデル毎に 5 回ずつ実行)。

4 評価実験

使用モデル 本実験では OpenAI の gpt-5.2[2]、gpt-5[3] 、gpt-5-mini[4] の 3 種類のモデルを使用する。推論労力 (reasoning effort) は全モデルで medium に設定する。

対象論文 コメンタリ論文生成の対象として、哲学者ダニエル・デネットの哲学方法論に関する筆者らの論文 [5] を用いた。筆者ら自身の論文を対象とすることで、生成論文を詳細に評価できる。

評価方法 生成されたコメンタリ論文を、gpt-5.2 を査読者モデルとして用いて評価する。評価時には、世界的に共有された哲学論文の評価基準を精緻化したものとして知られる “Philosophy Paper Grading Rubric”[5] を、査読者モデルのプロンプトに用いる。本ルーブリックは、(1) 論証の強さ・明確さ (45 点)、(2) 問題に対する理解やアイデアの独創性 (35 点)、(3) 自他の主張の統合 (5 点)、(4) 論文の構成 (15 点)、(5) 可読性 (5 点) の 5 つの評価カテゴリから構成され、各カテゴリの合計 (0–100 点) を評価スコアとする。また本実験では各モデルで 5 回ずつ実行し、生成結果のばらつきを検証する。

5 実験結果

本節では実験結果を説明する。生成原稿の一部を付録 B に示す。費用面では、gpt-5.2 を用いた場合、1 回あたり約 1 ドルでコメンタリ論文を生成できた。

図 2 に各モデルが生成した原稿の評価スコア分布 (各モデル 5 回) を示す。平均的には gpt-5.2 が最も

2) https://platform.openai.com/docs/models/gpt-5.2
3) https://platform.openai.com/docs/models/gpt-5
4) https://platform.openai.com/docs/models/gpt-5-mini
5) https://dailynous.com/wp-content/uploads/2017/05/lewin-philosophy-paper-grading-rubric-1-xl-page.pdf

表 1　生成されたアイデア戦略の分布。

モデル名	推論批判	前提・原理批判	含意の拡張	異論と応答
gpt-5-mini	4	6	5	10
gpt-5	9	7	2	7
gpt-5.2	16	8	1	0
総計	29	21	8	17

表 2　最終的に選択されたアイデア戦略の分布。

モデル名	推論批判	前提・原理批判	含意の拡張	異論と応答
gpt-5-mini	2	2	0	1
gpt-5	2	1	2	0
gpt-5.2	5	0	0	0
総計	9	3	2	1

高く、次いで gpt-5、gpt-5-mini の順であった。またアイデア生成・選択で用いられた戦略の分布を表 1 および表 2 に示す。生成されたアイデアでは「推論批判」が最も多く、gpt-5.2 は「異論と応答」を生成しなかった。また gpt-5.2 では、すべての試行で「推論批判」が選択され、選択アイデアも互いに類似していた。これらのアイデアはいずれも、対象論文の提案する立場が問題に十分対処できていない点を指摘する内容であった。

6　考察

6.1　実験結果の考察

評価スコアは gpt-5.2、gpt-5、gpt-5-mini の順に高かった（図 2）。査読者モデルに gpt-5.2 を用いることによる自己選好バイアスの可能性を考え、gpt-5 を査読者モデルとして再度評価したところ、同様に gpt-5.2 の生成原稿が平均して最も高く評価された。したがって本実験結果は、少なくとも自己選好バイアスのみでは説明されない。

筆者らによる定性的確認では、対象論文の要約は高品質であり、人間研究者の要約と区別しにくい水準であった。一方、批判的検討は論点の掘り下げが不十分な場合もあるが、議論はよく整理されており、対象論文の弱点を明確に指摘していた。今後は AI による生成物であることを伏せた上で、査読に近い条件で評価する。

また、アイデア生成・選択にはモデル差が見られた。gpt-5.2 では 5 回すべてで「推論批判」が選択され、選択アイデアも類似していたのに対し、gpt-5-mini では戦略と内容の多様性が相対的に高かった。この結果は、モデル選択がアイデアの多様性に影響しうることを示唆する。

6.2　倫理的考慮事項

本研究によって生じうる倫理的懸念を述べる。第一に、LLM によって生成された哲学論文が、人間によって書かれたと偽って投稿・出版されることには倫理的問題があると考える。こうした AI によって自動化された研究が哲学研究や研究コミュニティに及ぼす影響や、LLM を哲学論文の著者として良いか、また LLM によって生成された哲学論文をジャーナルに掲載すべきかについては、現在議論が進行中であり、今後も継続されるべきである [6]。

第二に、より哲学的な問題として、哲学という人間本性と深く関わる営みを自動化することは、人間本性に対する危機と受け止められる可能性がある [7]。しかし、哲学の論文生産が自動化されたとしても、各人の哲学的考察の自由が残されている限り、この危機は深刻ではないと考えられる。

7　関連研究

AI を用いた科学研究の自動化として、科学研究の各段階に対する補助と、End-to-End の全自動化手法がそれぞれ提案されている [1]。全自動化の代表例として、The AI Scientist [4] および The AI Scientist-v2 [2] は機械学習研究の工程をほぼ一通り実行する。その他のより改善された手法 [8] や、研究能力を評価するベンチマーク [9] も提案されているが、いずれも機械学習分野での自動化研究であり、人文学分野を対象としていない。

哲学に関しても、執筆のエンハンスメント [10]、思考実験に対する AI の判断 [11, 5]、特定の哲学者の模倣 [12, 13] などの研究がなされているが、これらの研究は哲学研究の自動化を目指すものではない。

8　結論

本研究では、哲学論文に対するコメンタリ論文を、アイデア生成から原稿生成まで一貫して自動生成する手法として AI 哲学者を提案した。評価実験では、LLM による自動評価において gpt-5.2 が高いスコア（100 点中 85 点以上）を示し、筆者らの確認でも質の高いコメンタリ論文が生成されていた。今後の課題として、参考文献の自動収集と直接引用の精度向上、ならびに実際の査読に近い条件での評価実験が考えられる。

謝辞

本研究はサントリー文化財団 2024 年度若手研究者のためのチャレンジ研究助成および日本財団 HUMAI プログラムの助成を受けたものです。

参考文献

[1] Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. Ai4research: A survey of artificial intelligence for scientific research. **arXiv**, 2025.

[2] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. **arXiv**, 2025.

[3] Alan Hájek. Philosophical heuristics and philosophical methodology. In **The Oxford Handbook of Philosophical Methodology**. Oxford University Press, 05 2016.

[4] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. **arXiv**, 2024.

[5] Arata Matsuda and Masashi Takeshita. What should philosophers do with “deceptive” intuition pumps? restrictionism vs reformism. **Philosophical Psychology**, pp. 1–20, 2025.

[6] Nicholas Hadsell, Rich Eva, and Kyle Huitt. Publishing robots. **Inquiry**, pp. 1–27, 2025.

[7] Neil Levy. The work of philosophy in the age of mechanical reproduction: How ai threatens the meaningfulness of philosophy. **Ergo: An Open Access Journal of Philosophy**, forthcoming.

[8] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In **The Thirteenth International Conference on Learning Representations**, 2025.

[9] Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. MLR-bench: Evaluating AI agents on open-ended machine learning research. In **The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2025.

[10] Sebastian Porsdam Mann, Brian D Earp, Nikolaj Møller, Vynn Suren, and Julian Savulescu. Autogen and the ethics of co-creation with personalized llms—reply to the commentaries. **The American Journal of Bioethics**, Vol. 24, No. 3, pp. W6–W14, 2024.

[11] Kiichi Inarimori, Masashi Takeshita, Arata Matsuda, and Kengo Miyazono. Philosophical significance of artificial intuitions. **PsyArXiv**, June 2025.

[12] Eric Schwitzgebel, David Schwitzgebel, and Anna Strasser. Creating a large language model of a philosopher. **Mind & Language**, Vol. 39, No. 2, pp. 237–259, 2024.

[13] Paul Smart, Robert Clowes, and Andy Clark. Chatgpt, extended: large language models and the extended mind. **Synthese**, Vol. 205, No. 6, p. 242, 2025.

A　使用したプロンプトの概要

システムプロンプト

```
## Role
Act as an excellent professional analytic philosopher.

## Task
Your task is [...]

## Restrictions on Style
- Write in a plain, clear, analytic philosophy style.
- Do not write in a literary, poetic, or essayistic style.
- Do not use metaphors, rhetorical devices, or evocative language.

## Restrictions on Terminology
- Adhere strictly to the terminology of the target article.
- Do not introduce new labels, expressions, or formulations unless unavoidable.
- Use consistent terminology.
- Do not use different labels to refer to the same thing.
```

ユーザープロンプト

```
## Input
You will receive the target article [...]

### Target Article
Title: {title}
Abstract: {abstract}
Keywords: {keywords}
Full Text: {full_text}

### Task
[Describe the task here.]

---

## Output Format
[Describe the required output format here.]

Example structure:
[Provide an example JSON structure here.]
```

B　提案手法により生成された原稿の要旨および原稿の冒頭

Abstract: The target article rejects Dennett’s restrictionism about “deceptive”’ intuition pumps in favor of reformism, motivated by the problem of over-restriction once experimental philosophy is taken seriously. This commentary focuses on the target article’s own framing of deceptiveness as an authority problem: after “turning all the knobs,” absent a persuasive argument, competing variants have “the same status” and lack “demonstrated authority.” I argue that, as presented, the target article’s two candidate reform strategies—artificial intuitions and aggregating intuitions—do not yet supply the required persuasive argument for treating the prioritized intuition as having demonstrated authority. The point is not to reject reformism as a research program, but to clarify the justificatory burden reformism incurs on Dennett’s terms.

Introduction: The target article asks what philosophers should do with “deceptive”’ intuition pumps and answers by rejecting Dennett’s restrictionism in favor of reformism. It begins from the familiar methodological picture on which thought experiments (Dennett’s “intuition pumps”) elicit intuitions that function as evidence for or against philosophical theories, and it adopts Dennett’s worry that some such devices are “deceptive” insofar as small adjustments to a case description can elicit conflicting intuitions and hence conflicting conclusions. The paper’s argumentative structure is then two-part: first, it reconstructs Dennett as favoring restrictionism and argues that restrictionism faces a problem of over-restriction once experimental philosophy is taken seriously; second, it proposes reformism and offers two candidate reform strategies—artificial intuitions and aggregating intuitions—as ways to determine which of the conflicting intuitions should be prioritized while continuing philosophical inquiry. [...]