輪講 2023-10-10
M1 櫛田怜奈

Jimpei Hitsuwari, Yoshiyuki Ueda, Woojin Yun, Michio Nomura. (2023).

# Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry

*Computers in Human Behavior*, 139.

1. Introduction
   - ◆ 近年、自然言語処理・生成技術の飛躍的な向上により、AI は人間が創作したものに酷似した文学や詩を創作できるようになった。

   - ◆ AI アートを評価する先行研究では、アルゴリズムと人間が作ったアートを比較し、参加者に作者の判別（人間 vs. AI）と美的スコアの評価を求めている（ex. Chamberlain et al., 2018; Gangadharbatla, 2022; Ueda et al., 2021）。

   - ◆ Schwartz (2015)は自身のウェブサイトで、閲覧者に人間が作った詩とアルゴリズムで生成した詩を比較させたところ、アルゴリズムで生成した詩の 65%が作者を欺いた（つまり、AI が書いた詩が、人間が書いたように見えた）。
     - ➢ その後、Lau et al. (2018) は、言語、リズム、韻のモデルからなる Deep-speare によって 14 行のソネットを生成し、一般人と英文学の専門家に評価を依頼した。
       - ✧ 一般人は AI が生成したものと人間が作ったものを判別できなかったが、専門家は読みやすさと喚起される感情について、前者が後者より劣っていると判断した。なお、この研究では、専門家は 1 人しかいなかったため、他の専門家も判別できるのか、それともこの判別はこの特定の専門家だけのものなのかは不明である。
     - ➢ これらの研究では、AI が生成した詩は最終的に人間によって選ばれた。

   - ◆ Kobis and Mossink (2021)は、AI 詩への人間の関与をより厳密にするために、AI 詩を Human-out-of-the-loop (HOTL)と Human-in-the-loop (HITL)に区別し、前者は AI が生成した詩から項目を選択する際に人間の介入を必要とせず、後者は AI が生成した詩を人間が選択することを必要とした。

- その結果、HOTL を用いた AI の詩は人間が作ったものと判別できたが、HITL を用いたものは判別できなかった。つまり、人間の関与によって AI 詩の質を高め、人間と同等にすることができることが明らかになった。

- さらに最近、Gunser et al.（2022）は Kobis and Mossink（2021）をより多くの材料を用いて再現し、参加者は HITL を用いた人間の詩と AI の詩を判別できないことを明らかにした。同時に、人間の詩は、感動的、よく書かれているなど、複数の評価において優れていることがわかった。

◆ 本研究では、これまでの研究成果を世界一短い詩である俳句に拡張することを試みる。

- 日本発祥の俳句は、5〜7〜5 音節の定型や「季語」と呼ばれる季節の単語を含むなど、明確なルールを持っている（Iida, 2008）。
  こうした俳句のルールが、俳句を他の詩の形式とは異なるものにしている。
  - 第一の側面は曖昧さであり、俳句は文字数が少ないため表現が曖昧であり、読者は少ない情報から情景を補足・解釈しなければならない（Hitsuwari & Nomura, 2022a; 2022c）。
    したがって、AI が曖昧な俳句を生成しても、読者は自分が想像できる情景に従って解釈しようとする。
  - もうひとつの側面は、俳句の印象が想起される心象がひとつかふたつに依存することである（Blasko & Merski, 1998; Hitsuwari & Nomura, 2022b）。
    他のタイプの詩は一般に俳句よりも長く、場面によってさまざまなイメージに変換されることが多いため、場面間の一貫性や物語性を持たせるという問題が解消される。

◆ 本研究では、人間が作った俳句と AI が生成した俳句の美しさを評価する要因をさらに検討した。

- 先行研究では、いくつかの項目で詩を評価している（詳細は Table 1）。
  - 対象物の美の体験の一般的な特徴を明らかにするために、Brielmann et al.（2021）は、著名な美学の哲学者が考察してきた 11 の次元（快楽、体験を続けたい、生きていると感じる、体験が誰にとっても美しいと感じる、体験とのつながりを感じる数、憧れ、欲望から解放された感じ、心がさまよう、驚き、体験をもっと理解したい、体験が物語を語っていると感じる）と、
    心理学者が伝える 8 つの次元（複雑さ、覚醒や興奮、体験から学ぶ、理解したい、多様性の調和、意味深さ、期待以上、興味）を提唱した。

- ✧ 本研究では、これらの次元を用いて、人間が作った俳句と AI が生成した俳句における美の経験を決定づける要因を特定した。
- ➢ さらに、以前の俳句研究（Hitsuwari & Nomura, 2022c）でも関与している畏怖と郷愁の感情についても探索的に検討した。

◆ AI アートの評価において問題となる心理的特性は、アルゴリズム嫌悪、つまり美しいものは AI ではなく人間によって作られるという信念である（Burton et al.）。

- ➢ アルゴリズム嫌悪は AI の詩を好むかとも関連している（Kobis & Mossink, 2021）。
- ➢ Okanda et al.（2019）は、アニミズムや共感特性がロボットの道徳的側面の評価に影響を与えることを示したことから、これらの特性の個人差がロボットや AI に対する印象に影響を与え、人間が作った俳句と AI が作った俳句の判別に関係している可能性がある。

**Table 1**
Literature review comparing human-made and AI-generated poetry.

| ID | Study | HITL or HOTL | AI algorithm | Human poems | Number of evaluators | Number of poems | Rating | Discriminating | Limitations | Notes |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Schwartz (2015) | HITL | Various algorithms (e.g., RACTER & RKCP) | Professional poets (e.g., William Blake & Frank O'Hara) | Thousands of people | NA | NA | - 65% of them could not identify the author | - Not empirical study (without control of experimental conditions or participants) | - First attempt to examine whether human and AI poetry can be discriminated |
| 2 | Hopkins and Kiela (2017) | HITL | LSTM | Classic poets (e.g., Shakespeare) | 70 | Eight manmade and two AI-generated poems | - Human poems were rated slightly higher quality (in total score of form, readability, and emotional evocation) than AI-generated poems<br>- The poems judged to be the most human and aesthetic were AI-generated | - 48.6% of human poems were falsely attributed to AI and 53.8% of AI poems were falsely attributed to humans | - Small sample size of poems and participants | |
| 3 | Crowdworkers in Lau et al. (2018) | HOTL | Deep-speare | Professional poets (e.g., Shakespeare) | 1000 | 50 manmade and 180 AI-generated quatrains | NA | - Accuracy was 53.2%, indicating it was hard to distinguish between human and AI poems | - No statistical tests were performed | - Each participant rated a few poems |
| 4 | Expert in Lau et al. (2018) | HOTL | Deep-speare | Professional poets (e.g., Shakespeare) | 1 | 30 manmade and 90 AI-generated quatrains | - AI-generated poems was better in meter and rhyme than human-made.<br>- Human-made poems were better in emotion and readability than AI-generated | NA | - Only one evaluator | - Evaluator was an expert in English literature |
| 5 | Lc (2021) | HITL | GPT-2 | Author made | 25 | 5 manmade and 5 AI-generated poems | - Evaluation regarding structure was not significantly different, but expression was rated higher for human-made than AI-generated poems | - Accuracy was 61.6% for human-made poems and 56.0% for AI-generated poems, but there was no significant difference between them | - Small sample size of poems and participants | - the AI could reproduce the nuance of the original text |
| 6 | Study 1 in Köbis and Mossink (2021) | HITL | GPT-2 | Novice made | 192 (About half of them participated in the discrimination task) | 20 manmade and ten AI-generated poems | - Human-made poems were rated higher than AI-generated poems with a 57.0% probability. | - 50.2% of all answers correctly discriminated the author | - Only HITL | |
| 7 | Study 2 in Köbis and Mossink (2021) | HITL & HOTL | GPT-2 | Classic poets (Maya Angelou & Hermann Hesse) | 384 (185 participated in the discrimination task) | Ten manmade, ten AI-generated with HITL, and ten AI-generated with HOTL poems | - Participants chose the better of the human-made and AI-generated poems, and human-made poetry was higher evaluated in 64.9%than AI-generated. (62.9% for HITL and 66.9% for HOTL) | - 65.5% in the HOTL poems and 53.7% in the HITL poems could be discriminated | - Little variety in human-made poems | - First experiment comparing HOTL and HITL |
| 8 | Study 1 in Gunser et al. (2022) | HITL | GPT-2 | Experts ( Gunser et al., 2022) | 120 | 18 manmade and 18 AI-generated with HITL poems | - Human-made poems were rated higher in 5 evaluations (well-written, inspiring, fascinating, interesting, and aesthetic) than AI-generated | - 40.3% of human-made poems were falsely attributed to AI, and 42.0% of AI-generated poems were falsely attributed to human-made | - Only HITL | - Expert poets and GPT-2 created continuations of the classical poems |
| 9 | Study 2 in Gunser et al. (2022) | HITL | GPT-2 | Classic poets (Franz Kafka, Friedrich Hölderlin, Robert Gernhardt, & Paul Celan) | 302 | 18 manmade and 18 AI-generated with HITL poems | - Human-made poems were rated higher in five evaluations (well-written, inspiring, fascinating, interesting, and aesthetic) than AI-generated | - 33.5% of human-made poems were falsely attributed to AI, and 40.2% of AI-generated poems were falsely attributed to human-made | - Human-made poems could be easily detected due to their historical language usage and sublime style | |

2. Hypotheses

- 本研究は、俳句が AI、人間、そしてそれらのコラボレーションによって創作された文芸形式として、美しさスコアや作者の判別にどのような影響を与えるかを検証することを目的とした。
  - 主な論点は 4 つあり、最初の 2 つは作品の評価に焦点を当てたが、後の 2 つはより作者帰属に焦点を当てた。

- 第一に、人間が作った俳句、人間が介在せずに AI が作った俳句（すなわち HOTL）、人間が介在して AI が作った俳句（すなわち HITL）の美しさスコアについて比較した。
  - 人間の関与は AI 詩の質を高める可能性が高い（e.g., Kobis & Mossink, 2021）ため、HITL 俳句は HOTL 俳句よりも評価されると考えられる。
  - AI が生成した俳句は、少ない情報でも十分に想像力を働かせることができるため、人間が作った俳句と同じか、それ以上に評価される可能性がある。

- 第二に、人間が作った俳句と AI が作った（HOTL/HITL）俳句の美しさを説明できる要因について検討した。
  - 先行研究（Brielmann et al., 2021）に従い、美的評価は哲学的観点と心理学的観点の両方に関連すると予想した。

- 第三に、参加者が、人間が作った俳句と AI が作った俳句（HOTL/HITL）を判別できるかどうかを検証する。
  - 典型的な詩を用いた先行研究（Kobis & Mossink, 2021）を参照すると、参加者は HOTL と人間の俳句を判別できるが、HITL と人間の俳句を判別できないと予想される。
  - しかし、俳句は他の詩に比べて情報量が少なく、読者の想像力に依存する部分が大きいことから、HOTL 俳句であっても人間俳句と判別できない可能性がある。

- 最後に、作者判別の精度を説明する要因、特に参加者の背景（学歴や俳句経験）や性格特性（ロボットや AI に対する態度と関連するアニミズムや共感性）が何かを検証する。
  - 参加者が美術を専攻しているか否かによって作者の判別精度が変わらなかったという Chamberlain et al.（2018）の知見から、参加者のバックグラウンドは人間が作ったものと AI が作ったものの判別精度に大きな影響を与えないことが予想された。
  - アルゴリズム嫌悪は、美しい作品を即座に人間の作者であると判断するため、アルゴリズム嫌悪を減少させるアニミズムや共感は、作者判別に有利に働くと予想された。

3. Methods

3.1. Participants

◆ クラウドワークス（https://crowdworks.jp）を通じて 400 名の日本人参加者を募集し、ウェブ上で実験を行った。

- アンケートで特定の回答を求める注意力チェックで適切な回答が得られなかった 15 名の参加者は除外した。
  385 人の参加者（男性 191 人、女性 194 人、M age = 40.9、SD = 10.1）のデータが最終的な分析に含まれた。
- すべての参加者は、研究の目的、方法論、危険性、実験期間、参加の任意性、参加辞退の権利、参加者情報の取り扱い方について説明を受けた。さらに、実験に参加する前に、書面によるインフォームド・コンセントを得た。参加者には 500 円が支払われた。

3.2. Materials

3.2.1. Haiku stimulus

◆ 俳句の季節とジャンルを分散させるために、俳句に必ず含まれる季語を 10 個選んだ。それぞれの季語についてプロが詠んだ句が数句収録されている『歳時記』から、人が詠んだ 40 句（10 季語各 4 句）を選んだ。

- 歳時記に掲載されている俳句は、いずれもプロの評価が高い。
- 事前のアンケートでは、8 人の一般人（実際の参加者には含まれていない）がこれらの俳句を見たことがないことを確認した。

◆ AI が生成する俳句については、北海道大学調和系工学研究室が開発した LSTM（Long Short-Term Memory）アルゴリズムに基づく俳句生成システムを使用した（Kawamura et al., 2021; Yokoyama et al., 2019）。

- システムはまず、LSTM によって俳句データが学習された言語モデルを用いて、俳句候補文のセットを生成する。次に、生成された文章に形態素解析を適用し、季語定型俳句の形式と一致するものを選択する。生成された文と学習データの俳句との類似度を計算し、ある閾値よりも類似度の高い文は削除される。
- 季語の頻度によって、生成される俳句の数は異なる。10 種類の季語それぞれについて、システムは 36,442 から 624,130 の俳句を生成した。
- アルゴリズムでは、AI 自身が日本語としての妥当性を計算し、最終的に上位 500 句を AI が生成した俳句としている。

- AI が生成した俳句から無作為に 20 句（10 季語各 2 句）を選び、HOTL 俳句とした。
- 3 人のアマチュア／初心者の俳人によって、美しさが高く評価された 20 句（10 季語各 2 句）を HITL 俳句とした。

- 80 句を 40 句からなる 2 つの刺激リストに無作為に分け、各リストは 20 句の人生成俳句、10 句の HOTL 俳句、10 句の HITL 俳句から構成された。

3.2.2. Questionnaires to investigate personality traits

- 性格特性が判断に及ぼす影響を検討するために、5 種類の質問紙を用いて性格特性を測定した。

- まず、対人反応性尺度（IRI）（Davis, 1980; Himichi et al,. 2017）で共感特性を測定した。
  - この尺度は、共感的関心、個人的苦痛、視点取得、想像性の 4 つの下位尺度からなり、5 段階のリッカート尺度で評価する。

- 第二と第三に、アニミズムの特性を測定するために、成人用アニミズム尺度（Ikeuchi, 2010）とアライブネス・アニミズム尺度（Okanda et al., 2019）を用いた。
  - 前者は、自然物の神格化、所有者の分身化、所有物の擬人化の 3 つの下位尺度からなる 5 段階リッカート尺度の 11 項目尺度である。
    - この尺度は、生命を持たないにもかかわらず、無生物に神性や生命の存在を感じる傾向を反映している。
  - 後者の尺度は、無生物が生きていると思う傾向を反映したもので、8 つのアイテム（火のついたろうそく、電話、時計、人形、テディベア、バナナの皮、雲、泥）と 4 つの植物（木、花、草、野菜）から、生きていると思うものをすべて選んでもらう。

- 第四に、俳句や美術に関する知識や関心を測定するために、以下の質問を用いた。
  - 俳句経験（「小学校や中学校で習っただけ」、「大人になってから触れる機会が少しある」、「大人になってから触れる機会がある」、「大人になってから触れる機会が多い」、「大人になってから毎日触れる機会がある」）
  - 美術館に行く頻度（「月に 2 回以上」、「月に 1 回程度」、「年に数回」、「年に 1 回程度」、「ほとんど行かない」、「行ったことがない」）
  - クリエイティブな仕事の経験（「まったくない」、「少しある」、「とてもある」、「かなりある」）
  - 芸術への関心（「まったくない」、「ほとんどない」、「少しある」、「とてもある」、「か

なりある」）

- 第五に、探索的研究として、Sugimori & Kusumi（2014）が開発した 4 項目尺度を用いて、「既視体験の頻度」と「類似性への感受性」を測定した。
  - 2 項目はデジャヴ体験の頻度を 7 段階のリッカート尺度（1: 過去 1 年間体験しなかった ～ 7: 毎日）で測定し、残りの 2 項目はデジャヴ体験が自分に当てはまる度合いを 5 段階のリッカート尺度（1: 間違いなく当てはまらない ～ 5: 間違いなく当てはまる）で測定した。

3.3. Procedure

- 参加者はインターネット上の実験ページにアクセスした。まずインフォームド・コンセントを読み、同意の上で実験を開始した。
- 実験は、俳句を評価する「評価ブロック」、俳句の作者が人間かどうかを判断する「判別ブロック」、参加者の性格特性を測定する「特性ブロック」の 3 つのブロックで構成された。
- Chamberlain et al. (2018)に従い、作者帰属の影響（AI が制作した作品がリストに含まれているかどうかの事前知識）を、参加者の半数が最初に評価ブロックを完了し、次に判別ブロックを完了することでコントロールした。
  残りの半数の参加者は、最初に判別ブロックを完了し、次に評価ブロックを完了した。
  - 両セットの参加者とも、特性ブロックは実験の最後に完了した。
- さらに、両セットの参加者に 2 つの異なる俳句リスト（それぞれ 40 句を含む）を提示し、両方のリストにおける結果の一貫性を調査した。

評価ブロック

- 人間が作った俳句と AI が作った俳句が個別に提示された。
- 参加者は、哲学的要素を含む、21 の要素を用いて、7 段階のリッカート尺度で俳句を評価するよう求められた（詳細は Brielmann et al., 2021）。

判別ブロック

- 俳句が提示され、参加者はそれぞれが人間によって作られたものか、AI によって作られたものかを判断するよう求められた。
- すべての俳句を判断した後、その判断の手がかりを次の 12 項目から選択した。
  言葉のリズム、一貫性、規則性、反復性、複雑さ、深さ、抽象性、意図性、独自性、表現、ニュアンス、その他
- 最後に、俳句を人間が作ったものか AI が作ったものかの選択について、自由記述で説明するよう求められた。

特性ブロック

- 年齢、性別、学歴、国籍に加え、5 種類の特性アンケートに回答してもらった。
  - 学歴は次のように分類された。
    (1)中学、(2)高校、(3)短大・高専、(4)大学、(5)大学院

3.4. Data analysis

- 第一に、R（ver.4.1.0）の anovakun 関数（ver.4.8.6; Iseki, 2021）を用いて、各人の美しさスコアをグループごとに平均し、一要因分散分析（人生成俳句、HOTL 俳句、HITL 俳句）で比較した。さらに、探索的分析のため、美しさ以外の 20 の評価についても同様の分析を行った。

- 第二に、線形混合モデル（lmer パッケージを使用（Bates et al., 2015））により、美しさ以外の 20 の評価が美しさスコアを説明できるかどうかを検討した。
  - 385 人の個人と 40 の俳句からなる合計 15,400 の観測値は、階層データと反復測定データ（マルチレベル・モデルと呼ばれる）の検定力に関する文献（Arend & Schafer, 2019）のサンプル・サイズを満たしていた。
  - 従属変数は美しさスコア、独立変数は美しさと俳句の条件以外の 20 の評価、学歴はコントロール変数であった。
  - 参加者、俳句、課題の順序（評価ブロックが先か判別ブロックが先か）はランダム効果に入れられた。独立変数はクラスター内で中心化された。学歴は、予備知識が俳句の評価に大きく影響する可能性があるため（Sato, 2007）、コントロール変数として含めた。

- 第三に、参加者が人間の俳句と AI の俳句を判別できるかどうかを調べるために、ヒット率が偶然水準（0.5）と有意に異なるかどうかを t 検定した。

- 最後に、個人の特性が判別精度を説明できるかどうかを調べるために、当たるか否かを従属変数、個人の特性を独立変数、参加者と俳句をランダム効果として一般化線形混合モデル（ロジスティック分析）で分析した。独立変数は全体平均を用いて中心化した。

4. Results

4.1. Beauty scores for human-made and AI-generated haiku

- Table 2 は、俳句評価の各項目についての記述統計量である。
- 人間条件、HOTL 条件、HITL 条件の美しさスコア（表 2 の一行目）を比較した。

- ➢ 多重比較の結果、HITL の得点が最も高かった。
- ➢ 人間条件と HITL 条件では得点に差はなかった（Figure 1 参照）。

◆ また、Table 2 は、その他の 20 の評価についての探索的分析の結果でもある。これらの結果から、AI が生成した俳句は、人間が AI の出力に介入（=HITL）した場合、人間が生成した俳句よりも高く評価されることが示唆された。

**Table 2**
Descriptive statistics for rating items of Haiku evaluation.

| | Mean (SD) | | | F value | p value | η^2 | Multiple comparison |
|---|---|---|---|---|---|---|---|
| | Human | HOTL | HITL | | | | |
| Beauty | 4.15 (.75) | 4.14 (.78) | 4.56 (.74) | 212.41 | .00 | .06 | Human = HOTL < HITL |
| Déjà vu | 3.30 (1.06) | 3.17 (1.05) | 3.50 (1.09) | 106.63 | .00 | .02 | HOTL < human < HITL |
| Image | 4.60 (.68) | 3.99 (.79) | 4.63 (.75) | 302.23 | .00 | .13 | HOTL < human = HITL |
| Valence | 4.11 (.58) | 4.00 (.60) | 4.28 (.64) | 92.19 | .00 | .04 | HOTL < human < HITL |
| Arousal | 3.96 (.69) | 3.82 (.74) | 4.08 (.74) | 75.88 | .00 | .02 | HOTL < human < HITL |
| Awe | 3.39 (.99) | 3.36 (1.02) | 3.68 (1.03) | 110.72 | .00 | .02 | Human = HOTL < HITL |
| Nostalgia | 4.14 (.89) | 4.08 (.96) | 4.28 (.90) | 42.19 | .00 | .01 | HOTL < human < HITL |
| Novelty | 3.63 (.70) | 3.57 (.73) | 3.64 (.77) | 6.70 | .00 | .00 | HOTL < human = HITL |
| Empathy | 3.81 (.87) | 3.53 (.92) | 3.97 (.89) | 181.03 | .00 | .04 | HOTL < human < HITL |
| Intention | 4.48 (.76) | 4.35 (.81) | 4.59 (.78) | 54.67 | .00 | .02 | HOTL < human < HITL |
| Joy | 3.40 (.83) | 3.31 (.87) | 3.57 (.87) | 68.85 | .00 | .02 | HOTL < human < HITL |
| Continue | 3.80 (1.00) | 3.71 (1.05) | 4.01 (1.00) | 92.46 | .00 | .02 | HOTL < human < HITL |
| Alive | 3.58 (.99) | 3.44 (1.02) | 3.73 (1.02) | 78.92 | .00 | .01 | HOTL < human < HITL |
| Universality | 3.82 (.77) | 3.79 (.80) | 4.17 (.77) | 177.46 | .00 | .05 | Human = HOTL < HITL |
| Longing | 3.58 (.97) | 3.49 (1.04) | 3.71 (.99) | 46.68 | .00 | .01 | HOTL < human < HITL |
| Free desire | 3.34 (.92) | 3.43 (.95) | 3.53 (.98) | 29.93 | .00 | .01 | Human < HOTL < HITL |
| Mind wandering | 3.20 (1.00) | 3.20 (1.02) | 3.39 (1.02) | 50.12 | .00 | .01 | Human = HOTL < HITL |
| Connection | 3.34 (1.02) | 3.31 (1.05) | 3.47 (1.03) | 31.06 | .00 | .00 | Human = HOTL < HITL |
| Surprise | 3.06 (1.04) | 2.95 (1.04) | 3.07 (1.06) | 22.74 | .00 | .00 | HOTL < human = HITL |
| Understand | 4.10 (.97) | 4.15 (1.03) | 4.27 (1.00) | 28.67 | .00 | .01 | HOTL < human < HITL |
| Tells story | 4.44 (.83) | 4.41 (.86) | 4.60 (.84) | 44.50 | .00 | .01 | Human = HOTL < HITL |

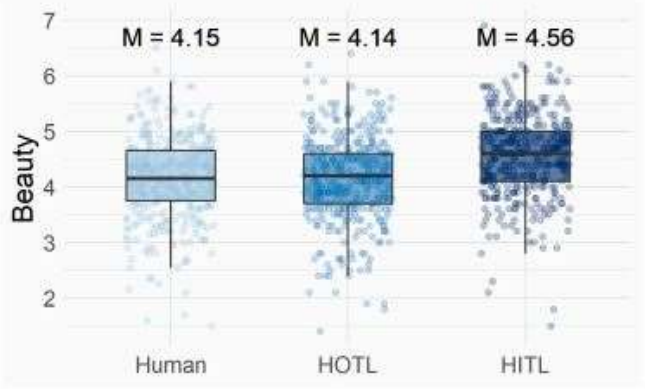


**Fig. 1.** Beauty scores in the human, HOTL, and HITL conditions.

4.2. Explanatory factors for beauty of human-made and AI-generated haiku

- ◆ 美的評価、課題の順番、俳句リスト、条件に関する 20 の因子が俳句の美しさを説明できるかどうかを線形混合モデルで検討し、美的評価の説明因子を最も多く決定した（Table 3）。
  - ➢ 美的評価に関連する因子のうち、生きている（b = - 0.01, SE = 0.01, t = - 0.56, p = .57）、心の迷い（b = 0.01, SE = 0.01, t = 1.17, p = .24）、つながり（b = - 0.01, SE = 0.01, t = - 1.31, p = .19）を除くすべての因子が俳句の美しさに寄与していた。
  - ➢ 課題順序（b = - 0.04, SE = 0.08, t = - 0.42, p = .67）、俳句リスト（b = - 0.06, SE = 0.06, t = - 1.04, p = .30）、人間と HOTL の条件差（b = 0. 11, SE = 0.09, t = 1.35, p = .18）は俳句の美しさを説明しなかったが、人間と HITL の条件差（b = 0.18, SE = 0.09, t = 2.16, p = .03）は説明した。
    - ✧ これらの結果から、4.1 節で示したように、人間が介在する AI 生成俳句（＝HITL）は、人間が作った俳句よりも美しさのスコアが高く、人間が介在しない AI 生成俳句（＝HOTL）のスコアは、人間が作った俳句と互換性があることが示された。
      また、俳句の美しさの評価には、タスクの順序（AI による作品かどうかの事前知識）は影響しないことが示された。
      さらに、この結果は、参加者に提示された刺激セットに依存しなかった。

**Table 3**
Results of the linear mixed model for factors explaining beauty.

| Random effects | Name | | Variance | | SD |
|---|---|---|---|---|---|
| ID | (Intercept) | | .49 | | .70 |
| Haiku ID | (Intercept) | | .09 | | .30 |
| Residual | | | .70 | | .84 |
| **Fixed effects** | **Estimate** | **SE** | **df** | **t value** | **p value** |
| (Intercept) | 4.16 | .07 | 227.90 | 62.36 | .00*** |
| Déjà vu | .09 | .01 | 14940.00 | 13.15 | .00*** |
| Image | .11 | .01 | 14990.00 | 16.26 | .00*** |
| Valence | .05 | .01 | 14970.00 | 6.32 | .00*** |
| Arousal | .04 | .01 | 14940.00 | 5.05 | .00*** |
| Awe | .10 | .01 | 14990.00 | 13.58 | .00*** |
| Nostalgia | .03 | .01 | 14960.00 | 4.19 | .00*** |
| Novelty | −.02 | .01 | 14940.00 | −3.24 | .00** |
| Empathy | .05 | .01 | 14950.00 | 6.22 | .00*** |
| Intention | .03 | .01 | 14950.00 | 3.82 | .00*** |
| Joy | .03 | .01 | 14980.00 | 3.81 | .00*** |
| Continue | .16 | .01 | 14930.00 | 16.83 | .00*** |
| Alive | −.01 | .01 | 14940.00 | −.56 | .57 |
| Universality | .28 | .01 | 14990.00 | 32.85 | .00*** |
| Longing | −.03 | .01 | 14930.00 | −3.32 | .00*** |
| Free desire | .01 | .01 | 14970.00 | 2.23 | .03* |
| Mind wandering | .01 | .01 | 14930.00 | 1.17 | .24 |
| Connection | −.01 | .01 | 14920.00 | −1.31 | .19 |
| Surprise | −.02 | .01 | 14950.00 | −2.40 | .02* |
| Understand | .05 | .01 | 14920.00 | 6.26 | .00*** |
| Tells story | .03 | .01 | 14930.00 | 3.22 | .00** |
| Task order | −.01 | .11 | 336.50 | −.08 | .94 |
| Haiku list | −.03 | .10 | 381.00 | −.25 | .81 |
| *Human vs. HOTL* | .11 | .08 | 75.18 | 1.35 | .18 |
| *Human vs. HITL* | .18 | .08 | 75.08 | 2.14 | .04* |
| Education | .04 | .04 | 381.00 | 1.12 | .26 |

Note. Dummy variables were set as follows: for Task order, −0.5 for the rating first condition and 0.5 for the discriminating first condition; for Haiku list, −0.5 for list 1 and 0.5 for list 2; for Human vs. HOTL, 1 for the HOTL condition; for Human vs. HITL, the HITL condition is 1, and the other two conditions are 0.

4.3. Distinguishing between human-made and AI-generated haiku

- 俳句の作者帰属能力（人間が作った俳句と AI が作った俳句を判別する能力）については、人間条件、HOTL 条件、HITL 条件のヒット率（俳句の作者を正しく判断する確率）は、それぞれ .55、.50、.43 であった（Figure 2）。
  - 人間条件でのヒット率は偶然水準より有意に高かった（$t(39) = 3.51$, $p = .001$）が、HITL 条件では偶然水準より有意に低かった（$t(19) = - 3.19$, $p = .005$）。HOTL 条件でのヒット率は偶然と差がなかった（$t(19) = 0.03$, $p = .98$）。
    - これらの結果は、参加者が HOTL 俳句について、人間が作ったものか AI が作ったものかを判別できなかったことを示唆している。

さらに、彼らは HITL 俳句の作者は AI ではなく人間であると考えていた。

◆ 美しいものは AI ではなく人間が作ったものであると考える（＝アルゴリズム嫌悪）のであれば、ヒット率と美しさスコアには関係があるのかもしれない。

- 各俳句の美しさスコアとヒット率の関係を探索的に調べた結果、人間条件では正の相関が見られたが有意ではなく（$r(39) = 0.18, p = .26$）、HOTL 条件と HITL 条件ではそれぞれ有意に負の相関が見られた（$r(19) = -.54, p = .01$ と $r(19) = -.47, p = .04$）（Figure 3）。
- AI 俳句の美しさスコアが高いほどヒット率が低いのは、参加者が「美しい俳句ほど人間が作った可能性が高い」と考えていることを反映している。

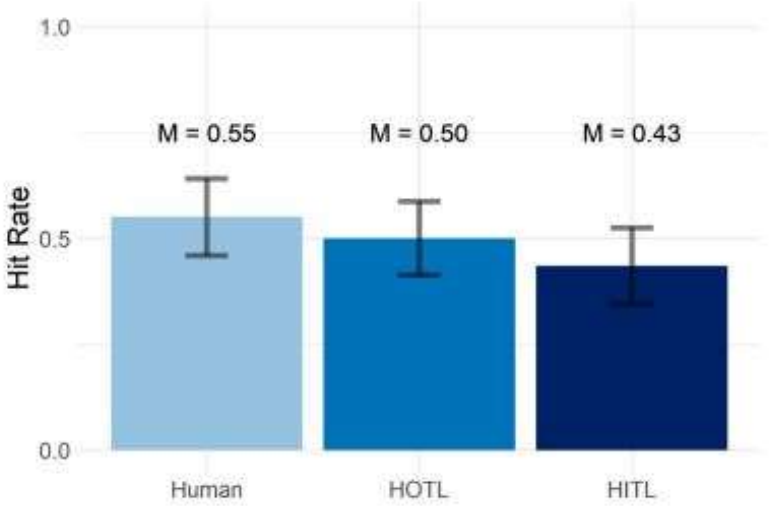


**Fig. 2.** Hit rate in the human, HOTL, and HITL conditions.

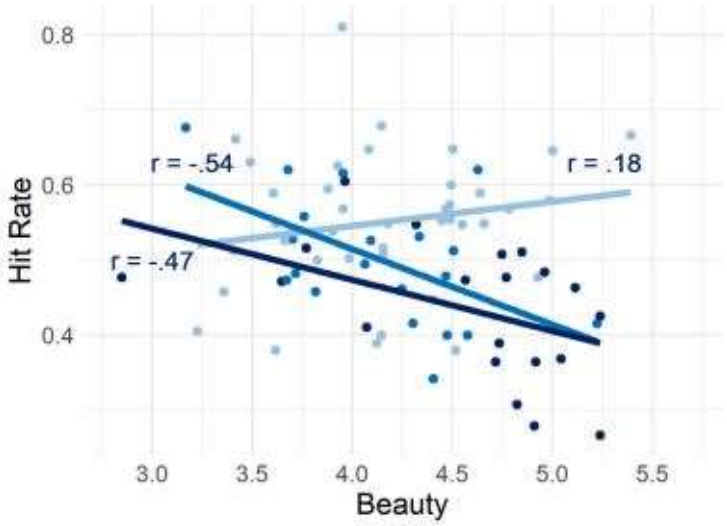


**Fig. 3.** The scatterplot between the beauty score and the hit rate.

4.4. The relationship between hit rate and personality traits

- 最後に、作者判別の精度を説明する要因について検討した。
  - これらの要因には、学歴、俳句経験、アニミズムや共感性などの性格特性が含まれる。
  - Table 4 に示すように、アニミズムの擬人化傾向が高いほど（b = 0.10, SE = 0.03, z = 2.81, p = .01）、また、俳句経験が多いほど（b = 0.07, SE = 0.03, z = 2.08, p = .04）、正答率が高かった。このことは、作品の特性だけでなく、擬人化傾向や経験といった参加者の要因が、俳句作者の判別能力に影響を与えていることを示している。

**Table 4**
Trait factors explaining the discrimination accuracy.

| Random effects | Name | | Variance | SD |
|---|---|---|---|---|
| ID | (Intercept) | | .02 | .16 |
| haikuID | (Intercept) | | .15 | .39 |
| **Fixed effects** | **Estimate** | **SE** | **z value** | **p value** |
| (Intercept) | .04 | .05 | .75 | .45 |
| IRI | | | | |
| Personal Distress | .06 | .03 | 1.88 | .06 |
| Empathic Concern | −.11 | .06 | −1.82 | .07 |
| Perspective Taking | .01 | .03 | .26 | .79 |
| Fantasy | .04 | .03 | 1.35 | .18 |
| Animism | | | | |
| Deification | .01 | .02 | .33 | .74 |
| Personalification | −.06 | .03 | −1.77 | .08 |
| Anthropomorphization | .10 | .03 | 2.81 | .01** |
| Aliveness Animism | .01 | .01 | .57 | .57 |
| Haiku Experience | .07 | .03 | 2.08 | .04* |
| Educational Background | .01 | .02 | .42 | .67 |

N: 15400, ID: 385, haikuID: 80.

4.5. Exploratory analysis: the rationale behind author discrimination

- 作者判別の根拠（参加者がどのような手がかりを重視して俳句の作者を判別したか）とヒット率との関係を調べた。
  - 人が作った俳句と判断する手がかりとして最も多かったのは「作品の深み」（参加者の 72%）であり、AI が作った俳句と判断する手がかりは「表現力」（参加者の 58%）であった。
  - 各参加者のヒット率を従属変数とする重回帰分析の結果、一貫性（b = 0.07, SE = 0.02, t = 3.91, $\beta$ = 0.21, p = .00）は人生成俳句のヒット率を説明でき、規則性（b = 0. 04、SE=0.02、t=2.24、$\beta$ =0.12、p=.03）、意図性（b=0.09、SE=0.02、t=4.10、

$\beta$=0.21、p=.00)、その他（b=-0.15、SE=0.04、t=-3.39、$\beta$=-0.17、p=.00）は AI が生成した俳句のヒット率を説明できた（Table 5）。

**Table 5**
Rationale behind author discrimination and hit rate.

| | Human | | | | | | AI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Estimate | SE | t value | Beta | p value | Mean | Estimate | SE | t value | Beta | p value |
| (Intercept) | | .56 | .02 | 29.64 | | .00*** | | .46 | .02 | 24.23 | | .00*** |
| Rhythm | .30 | .02 | .02 | 1.41 | .07 | .16 | .28 | −.02 | .02 | −.98 | −.05 | .33 |
| Consistency | .25 | .07 | .02 | 3.91 | .21 | .00*** | .41 | .00 | .02 | .26 | .01 | .79 |
| Regularity | .06 | .03 | .03 | .96 | .05 | .34 | .38 | .04 | .02 | 2.24 | .12 | .03* |
| Repeatability | .02 | .09 | .06 | 1.57 | .08 | .12 | .08 | .01 | .03 | .26 | .01 | .80 |
| Complexity | .30 | −.03 | .02 | −1.62 | −.08 | .11 | .13 | .02 | .02 | .70 | .03 | .49 |
| Depth | .72 | −.03 | .02 | −1.90 | −.10 | .06 | .23 | −.02 | .02 | −.97 | −.05 | .33 |
| Abstraction | .25 | .01 | .02 | .53 | .03 | .60 | .17 | .01 | .02 | .51 | .03 | .61 |
| Intentionality | .52 | .00 | .01 | .11 | .01 | .91 | .14 | .09 | .02 | 4.10 | .21 | .00*** |
| Uniqueness | .38 | .00 | .01 | .08 | .00 | .94 | .05 | −.05 | .03 | −1.58 | −.08 | .11 |
| Expression | .48 | −.01 | .01 | −.98 | −.05 | .33 | .58 | −.03 | .02 | −1.65 | −.08 | .10 |
| Nuance | .46 | −.01 | .01 | −.61 | −.03 | .54 | .36 | .01 | .02 | .68 | .03 | .50 |
| Others | .03 | .07 | .05 | 1.60 | .08 | .11 | .03 | −.15 | .04 | −3.39 | −.17 | .00*** |

## 5. Discussion

- 本研究では、明確なルールがあり、限られた文字数で表現される（＝限られた情報しか伝わらない）俳句について、人間と AI が作成した俳句を様々な観点から評価した。
  - 人間が選んだ AI の俳句は、人間が作った俳句よりも美しさスコアが高いことがわかった。また、AI がランダムに選んだ俳句は、人間が作った俳句と同程度の美しさスコアであった。
    - これらの結果は、人間と AI の協働がより優れた創造性につながること、また、少なくとも俳句制作においては、AI の生成能力が創造的分野において人間に匹敵することを示唆している。
- 俳句の美意識を説明する要因として、哲学的な美意識と心理学的な美意識があることも明らかになった。
- 作者の判別については、参加者は人間が作った俳句と AI が作った俳句を判別できなかった。この結果と前述した美しさの得点は、AI が人間と同等の品質の作品を作ることができることを示している。
  - さらに、擬人化傾向という性格特性や俳句経験が、人間が作った俳句と AI が作った俳句の判別能力に寄与している可能性がある。

### 5.1. The beauty score between AI-generated and human-made haiku

- 人間が詠んだ俳句と、人間が介在しない AI が詠んだ俳句の美しさスコアは同程度であった。
  - この結果は、人間が作成した俳句が最も好まれたという以前の研究（Kobis & Mossink, 2021）と矛盾している。

- ✧ この説明のひとつは、詩のスタイルの違いかもしれない。俳句は世界で最も短い詩の形式であり、一定のルールに従って情報を集約する必要があるため、AI 芸術の質は十分に確保されている。
- ✧ また、AI の訓練教材が、おそらく以前の研究で使用されたものよりも優れていたことにも注目すべきである。俳句は、5-7-5 音節の形式や季語を含めるといった明確なルールが特徴的であり、訓練が容易であった可能性がある。
- ✧ さらに、先行研究（Kobis & Mossink, 2021）では、AI が作成した詩の最初の 2 行を人間が作成した詩と同じにし、両方の詩を一緒に並べ、参加者にどちらを好むかを尋ねた（すなわち、二者択一強制選択）ため、その結果との不一致は、方法論の違いによって説明できるかもしれない。
  二者択一は、本研究で採用した絶対評価に比べ、2 つの詩の微妙な違いに比較的敏感である。本研究では、わずかな違いがある場合、非専門家にとって人間が作った作品の優位性が消失することを示した。

5.2. Factors explaining beauty

- ◆ 本研究では、人間が作った俳句と AI が作った俳句の美しさとヒット率を説明するいくつかの要因を特定した。
  - ➢ 例えば、人間の俳句の美しさを説明することが知られている心象の鮮明さ、肯定的な感情、畏怖や懐かしさといった高次の感情（Hitsuwari & Nomura, 2022b; 2022c）は、AI 俳句においても美しさの説明因子となることがわかった。
- ◆ さらに、美の実証的研究が始まるはるか以前から、哲学者や思想家が美を説明するために考えてきた哲学的要因（Brielmann et al., 2021）も詩の美しさに関連することが判明した。

5.3. Author discrimination of AI-generated and human-made haiku

- ◆ 人間が作った詩と AI が生成した詩を判別できなかったことは、これまでの知見と一致している（Hopkins & Kiela, 2017; Lau et al., 2018; Schwartz, 2015）。
  - ➢ 判別できなかった背景には、俳句の生成技術が向上していることが大きな要因として考えられる（Ito et al., 2018; Kawamura et al., 2021; Yokoyama et al., 2019）。
  - ➢ また、俳句は一般的な詩に比べて非常に短く、イメージに依存するなどの特徴があるため、人間が介在しない AI が生成した俳句でも判別に支障をきたしている可能性がある。

- ◆ AI が生成した俳句における美しさスコアとヒット率の間の負の関係を考慮すると、人々はアルゴリズム嫌悪を持っている可能性がある（Burton et al.）。
  - ➢ アルゴリズム嫌悪は、人間と AI の詩の比較という研究文脈において重要な概念で

ある（Kobis & Mossink, 2021）。
ここでは、人間が美しい俳句を作った、つまり AI 俳句を見下したという考え方が成り立つかもしれない。

5.4. Effects of experience, personality, and clues on author discrimination

- 擬人化傾向や共感的関心といった性格特性のような因子は、AI とロボットの相互作用研究において検討されてきた（e.g., Darling et al, 2015; Okanda et al, 2019）。
  - 本研究では、それらが俳句作品における作者の判別においても意味を持つことを示した。
  - 擬人化傾向の高い人は、美しい俳句を衝動的に人の手によるものだと判断せず、結果としてヒット率が高いことが推察される。

- 俳句に使用されている単語が互いに一貫しているかどうかという一貫性に注目した参加者は、AI が生成した俳句よりも人間が生成した俳句のヒット率が高いことがわかった。
- 対して、規則性や意図性に注目した参加者は、AI が生成した俳句のヒット率が高かった。AI が生成した俳句は、時折、意味が一貫せず、理解しにくいものがあった。
  - 一貫性に注目することは、人間が作った俳句を判別するためのツールとして機能した。
  - AI アートの鑑賞では規則性がしばしば認識され（cf. Chamberlain et al., 2018）、意図性の存在（または不在）は AI アート研究の主要なトピックの 1 つである（Chamberlain et al., 2018; McCormack et al., 2019）。

5.5. Limitations and future research

- 本研究では、人間が作成した俳句と AI が作成した俳句について、評価と判別の観点から検討し、美しさスコアやヒット率と様々な変数との関係を明らかにした。
- しかし、いくつかの限界があった。その一つは、HITL 条件における AI 俳句の選択方法である。
  - 本研究では、俳句の選定は 3 人の非専門家によって行われた。参加者のほとんどが非専門家であったため、専門家に選ばれた人生成俳句よりも HITL 俳句の方が理解しやすかった可能性がある。
  - 今後、非専門家が選んだ人生成俳句や、専門家が介在した HITL 俳句を用いた実験を行うことも考えられる。同時に、初心者や俳句のプロが参加者として美しさを評価することや作者を判別することも可能である。
  - とはいえ、少なくとも素人の視点に立てば、本研究は、人間と AI とのコラボレーションが優れた作品を生み出す可能性があることを示している。

- 本研究では、日本人の参加者には日本語の俳句のみを提示した。
  - 俳句生成の AI モデルは、他国でも洗練されつつあり（Hreskova & Machova, 2018; Wong, Chun, Li, Chen, & Xu, 2008）、俳句の東西文化比較では、美しさスコアに有意差はなかった（Hitsuwari & Nomura, 2022c）。
    この研究は他の文化にも適用できるかもしれない。

5.6. Conclusions

- AI が作成した俳句と人間が作成した俳句の美的評価と作者判別が行われた。
  - 美しさスコアは、人間が作った俳句と無作為に選んだ AI が作った俳句で同じだった。
  - しかし、AI が作った俳句であっても、人間が選んだものであれば、美しさスコアが最も高くなった。
  - 人間が作った俳句と AI が作った俳句の判別については、参加者が判別することは困難であった。
  - AI 俳句では、美しさスコアと判別能力の間に負の相関が見られた。
- これらの結果は、（情報が最小限である）俳句において、AI の芸術の質が人間に匹敵するレベルに達しており、人間と AI の協働によって、より創造的な作品を生み出すことができることを示唆している（Booten & Gero, 2021）。