06/21/2023 D2 服部エリーン彩矢

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (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*, 107502.
https://doi.org/10.1016/J.CHB.2022.107502

概要

- 人が生成した俳句と AI が生成した俳句，人が介在した AI 生成俳句を比較し，AI アートに対する美的評価と人々の信念について検討
  - 俳句は，美しさなどの 21 項目で評価
  - 人生成の俳句と AI 生成の俳句を判別

1. Introduction

- 人工知能（AI）アートが盛んになっている
  - その主なアート対象は，絵画などの視覚芸術だった(Elgammal et al., 2017; Gatys et al., 2016; Karayev et al., 2013; Li & Wand, 2016, see Cetinic & She, 2022 for review)
  - 近年は，自然言語の処理・生成技術が飛躍的に向上し，AI の創作する文学や詩は人の創作に酷似してきている
  - こうした AI アートは，生成側を議論するコンピュータサイエンスやロボット工学の分野の研究者や，美や創造性の理解を中心に議論する心理学や哲学の分野の研究者に注目されている (c.f., Daniele & Song, 2019)

- AI アートを評価するこれまでの研究では，アルゴリズムから制作されたものと人が制作したアートを比較し，参加者に製作者（人 vs. AI）の判別を求めている(e.g., Chamberlain et al., 2018; Gangadharbatla, 2022; Ueda et al., 2021)
  - 製作者の判別は，チューリングテスト（人が行う知的活動と機械や AI の能力が判別できないことを検証する）を AI アートに適用したもの
  - 絵画に関しては，人が制作したものと AI が制作したものが判別し難いという結果で一貫する (Chamberlain et al., 2018; Elgammal et al., 2017; Gangadharbatla, 2022; Ragot, Martin, & Cojean, 2020; Ueda et al., 2021)
    - 興味深いことに，好みや美しさの評価スコアには差が見られる(Ragot et al., 2020; Ueda et al., 2021, しかし Cham-berlain et al., 2018 では見出されなかった)
    - アートの好みは，実際の製作者ではなく，誰が制作者だと信じているかに影響されるという研究もある(i.e., 製作者帰属 Cham-berlain et al., 2018; Ragot et al., 2020,しかし Hong & Curran, 2019 では認められなかった)

- 近年の AI は文学や詩の生成ができるようになったことを踏まえると，言語芸術にまで議論を広げる価値があると思われる
  - 人が生成した文学や詩と AI が生成した文学や詩を比較する研究は，2010 年代後半から検討

されている(Gunser et al., 2022; Hopkins & Kiela, 2017; K ¨obis & Mossink, 2021; Lau et al., 2018; Lc, 2021; Oliveira, 2009; Schwartz, 2015; 概要は **Table 1** 参照)

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 - 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. - 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	- 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

(*continued on next page*)

Table 1 (*continued*)

ID	Study	HITL or HOTL	AI algorithm	Human poems	Number of evaluators	Number of poems	Rating	Discriminating	Limitations	Notes
							higher evaluated in 64.9%than AI-generated. (62.9% for HITL and 66.9% for HOTL)			
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	

- ✧ 回答者の **65%**が，実際はアルゴリズムで生成された詩を人が書いた詩と判断した**(Schwartz, 2015)**
- ✧ 短歌の評価において，一般人は，**AI** 生成と人生成の短歌を判別できなかったが，専門家は，**AI** 生成の短歌が読みやすさや喚起される感情に関して劣ると判断した**(Lau et al., 2018)**
  - ➢ 専門家がたったの **1** 名だったため，結論付けられない
- ✧ **K ¨obis and Mossink (2021)**は，上記 **2** つの研究に対して，人が **AI** 生成の詩に関与する影響を厳密にするため，参加者に提示する **AI** の詩を人が選択する場合**(human-in-the-loop; HITL)**と，人が選択に関わらない場合**(human-out-of-the-loop; HOTL)**を比較した
  - ● **HOTL** では，人生成と **AI** 生成の詩を判別できたが，**HITL** では判別できなかった
- ✧ 上記の結果は，より多くの刺激を用いて再現され，人生成の詩の方が「感動的」，「良く書けている」などの複数の評価で優れていることが明らかとなった**(Gunser et al., 2022)**

- 本研究では，これまでの知見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」と呼ばれる季節を入れるなど，明確なルールがある**(Iida, 2008)**

※松尾芭蕉の俳句を例に　**閑(しずか)さや　岩(いわ)にしみ入(い)る　蝉(せみ)の声(こえ)**
1689 年 **7** 月 **13** 日に，山形市の立石寺(りっしゃくじ)で詠んだ句。「蝉」が夏を表す季語となっている。

- ✧ 上記のルールから，俳句は他の詩の形式と異なる特徴を持つ
  - ● 曖昧さ：文字数が少ないため，読者は少ない情報から場面を補足・解釈しなければならない**(Hitsuwari & Nomura, 2022a; 2022c)**
    - ➡ **AI** が曖昧な俳句を生成しても，読者は自己解釈できる

- 呼び起される心象が **1** つか **2** つに依存する
  - ➡ 場面間の一貫性や物語性を持つ困難さがない
- ✧ これらの特徴から，**AI** が生成した俳句は人が生成したものと判別することが難しいと考えられる

- 本研究では，人が生成した俳句と **AI** が生成した俳句の美しさの評価に寄与する要因についてもさらに検討した
  - ➢ 先行研究で用いられた評価の項目数は，あまりに少ないために，俳句における美的感覚の喚起要因を検討できなかった
    - ✧ 例）**Hopkins & Kiela, 2017** では「読みやすさ」と「感情喚起」，**Gunser et al., 2022** では「よく書けている」，「感動的」，「魅力的」，「興味深い」，「美的」など；詳細は **Table 1** 参照
  - ➢ 本研究は，**Brielmann et al. (2021)**に従い，オブジェクトの種類を超えた美の体験の一般的な特徴として，美学哲学者が考案した **11** の次元と，心理学者が考案した **8** の次元を用いる
    - ✧ **11** の次元(悦び，体験継続願望，生き生きとする，万人にとって美しいと感じられる，体験と多くの繋がりを感じる，憧れ，欲求から解放されたと感じる，心がさまよう，驚き，体験をもっと理解したいと思う，体験から物語が伝わってくるように感じる)
    - ✧ **8** の次元(複雑さ，覚醒や興奮，体験から学ぶ，理解したいと思う，多様なものとの調和，有意義，自分の想像を超える，興味深い)
  - ➢ これにより，人生成の俳句と **AI** 生成の俳句における，美の体験を規定する因子を特定した

過去の俳句研究(**Hitsuwari & Nomura, 2022c**)でも関与している畏怖やノスタルジア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，美しいものは **AI** ではなく人が作るものだという信念，アルゴリズム嫌悪である(**Burton et al., 2020**)
  - ➢ これは，**AI** 詩の好みとも関連する(**K ¨obis & Mossink, 2021**)
  - ➢ アニミズム[1]や共感の特性が，ロボットの道徳的側面の評価に影響することが示されていることから(**Okanda et al., 2019**)，こうした個人特性が人生成の俳句と **AI** 生成の俳句の判別に関係する可能性がある

2. Hypothesis

本研究の目的：**AI** と人，そしてそれらの共同作業によって生み出された文学芸術としての俳句が，美しさの評価や生成者の判別にどのような影響を与えるかを検証すること

＜俳句作品の評価＞

1. 人が作成した俳句，人の介入なしで **AI** が生成した俳句(**HOTL**)，人が介入して **AI** が生成した俳句(**HITL**)の美しさスコアを比較する
   - ➢ 人の関与は **AI** の詩の質を高めると考えられるので(**e.g., K ¨obis & Mossink, 2021**)， **HITL** 俳句は **HOTL**

[1] アニミズムとは，**万物に霊魂**が宿るとする信仰のこと。人間だけでなく，動物や植物，天体や自然現象にも生命や意思があると考える思想。

よりも評価されるだろう

- 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，人が生成した俳句と同じ，あるいはそれ以上に評価される可能性がある

2. 人が作成した俳句と AI が作成した俳句(HOTL/HITL)の美しさを説明できる要因について検討する
   - 人が制作した芸術と AI が制作した芸術の美的感覚に影響を与える要因は多様であるが，AI による詩の研究では，膨大な数の評価項目を用いて取り組んだ研究はない
   - 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と関連すると予測する

＜生成者帰属＞

3. 参加者が，人が生成した俳句と AI が生成した俳句(HOTL/HITL)を判別できるかを検証する
   - 典型的な詩を用いた先行研究(K ¨obis & Mossink, 2021)を参照すると，参加者は HOTL と人の俳句を判別できるが，HITL と人の俳句を判別できないと予測される
   - しかし，俳句が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ことを考えると，HOTL の俳句であっても人の俳句と判別できない可能性がある

4. どのような要因，特に参加者の背景(教育や俳句の経験)や性格特性(ロボや AI に対する態度と関連するアニミズムや共感性)が，生成者の判別制度を説明するかを検証する
   - 参加者が美術専攻かどうかで制作者の判別精度が変わらないとした Chamberlain et al. (2018)の知見から，参加者の背景は，人生成と AI 生成のものの判別精度に大きな影響を与えないことが予測された
   - アルゴリズム嫌悪は，美術作品を即座に人が制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，生成者の判別に有利に働くと予測された

3. Methods

3.1. Participants

- 効果量が小さいことを仮定すると，俳句の生成条件(人，HITL，HOTL)を比較するには，322 人の参加者が必要であった
  - そこで，クラウドワークスを通して 400 人の日本人参加者を募集する Web 実験を実施した
    - アンケート内で，特定の回答を求める注意チェックで適切な回答を得られなかった参加者を 15 名除外した
- 385 名(男性 191 名，女性 194 名，*M*age =40.9, *SD* =10.1)の参加者を分析に含めた

3.2. Materials

3.2.1. Haiku stimulus

＜人の俳句＞

- 俳句の季節や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 個選択した

- 季語ごとにプロが作った俳句が数句掲載されている「歳時記」から，人が作った俳句 **40** 句(季語 **10** 語につき各 **4** 句)を選択した
  - 歳時記の俳句は，どれもプロから高い評価を得ている
  - 事前アンケートで，**8** 名の一般人(参加者以外)から「見たことがない」との回答を得ている

＜**AI** の俳句＞

- **AI** が生成する俳句には，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用いた(Kawamura et al., 2021; Yokoyama et al., 2019)
  - 本システムは，まず **LSTM** によって俳句データで学習した言語モデルを用いて，俳句の候補文のセットが生成される
    - 次に，生成された文章に形態素解析を施し，季節の定型俳句の形式と一致する文章が選択される
    - これにより，**10** 種類の季語に対して，**36,442**～**624,130** 句の俳句が生成された
    - アルゴリズムでは，**AI** 自身が日本語としての妥当性の度合いを計算し，最終的に上位 **500** 句を生成した

HOTL 俳句：**AI** 生成の俳句から，ランダムに選んだ **20** 句(季語 **10** 語につき各 **2** 句)とした

HITL 俳句：**3** 人のアマチュア・初心者俳人から「美しい」と評価された **20** 句(季語 **10** 語につき各 **2** 句)とした

- 全 **80** 句を **40** 句からなる **2** つの刺激リストにランダムに分け，リストごとに **20** 句の人の俳句，**10** 句の **HOTL** 俳句，**10** 句の **HITL** 俳句で構成した

3.2.2. Questionnaires to investigate personality traits

- 性格特性の測定には，**5** 種類の質問紙を用いた

1. 共感性の測定に，**Interpersonal Reactivity Index (IRI)** (Davis, 1980; Himichi et al., 2017)を用いた
  - **5** 件法

2.＆3. アニミズム特性を測定するために，**the Adult Animism Scale** (Ikeuchi, 2010) ＆ **the Aliveness Animism Scale** (Okanda et al., 2019)を用いた
  - 前者の尺度は，生命を持たない無生物に神性や生命の存在を感じる傾向を反映する(**5** 件法)
  - 後者の尺度は，生物が生きていると考える傾向を反映する(**8** つの物，**4** つの植物から，生きていると思うものを全て選ぶ)

4. 俳句やアートの知識・関心を測定する小項目
  - 俳句経験，美術館の訪問頻度，創造的な仕事の経験，アートに対する興味などを問う項目

5. 探索的研究として，**Sugimori and Kusumi (2014)**の **4** 項目を用いて，デジャヴ体験の頻度と類似性への感度を測定した
  - デジャヴ体験の頻度は **7** 件法，デジャヴ体験が自分に当てはまる程度は **5** 件法で測定

3.3. Procedure

- 実験は，俳句を評価する評価ブロック，俳句の生成者が人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 つのブロックから構成されていた
  - Chamberlain et al. (2018)に従い，生成者帰属の影響(AI が生成した作品がリストに含まれるかどうかの事前知識)をコントロールするため，参加者の半数が最初に評価ブロックを行い，次に判別ブロックに取り組んだ
    - 残りの半数の参加者は，まず判別ブロックを行い，次に評価ブロックに取り組んだ
  - 両者とも，特性ブロックは実験の最後に取り組んだ
  - さらに，両方の参加者に 2 つの異なる俳句リスト(それぞれ 40 句)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

Rating block

- 評価ブロックでは，人が生成した俳句と AI が生成した俳句が個別に提示された
  - 参加者は，美的評価に関連する心理的要素(美しさ，感情価，覚醒，畏敬，感動，鮮烈，情熱，新しさ，懐かしさ，デジャヴ)と，哲学的要素(悦び，継続願望，生き生きとする，普遍性，体験との繋がり，憧れ，欲求からの解放，心がさまよう，驚き，理解欲求，物語伝達性)，の 21 次元について，7 件法で評価した（詳細は，Brielmann et al, 2021 を参照）

Discriminating block

- 判別ブロックでは，俳句が提示され，それぞれが人によって生み出されたものか，AI によって生み出されたものかを判断した
  - 全ての俳句を判断した後，判断の手がかりとなった観点について，12 項目から選択した
    - 「言葉のリズム」「一貫性」「規則性」「繰り返し性」「複雑さ」「深さ」「抽象度」「意図性」「独自性」「表現」「ニュアンス」
  - 最後に，俳句を人が生成したものか，AI が生成したものかを選択した判断について，自由記述で説明するよう求めた

Trait block

- 特性ブロックでは，年齢，性別，学歴，国籍に加え，5 種類の特性を測定する質問紙に回答した

3.4. Data analysis

＜俳句作品の評価＞

1. 「人が作成した俳句，人の介入なしで AI が生成した俳句(HOTL)，人が介入して AI が生成した俳句(HITL)の美しさスコアを比較する」
   - 個別の美しさスコアを群別に平均して，一要因分散分析(人の俳句，HOTL 俳句，HITL 俳句)で比較した
     - 探索的分析のため，美しさ以外の 20 の観点の評価についても，同様の分析を行った

2. 「人が作成した俳句と AI が作成した俳句(HOTL/HITL)の美しさを説明できる要因について検討する」
   - 線形混合モデル(Bates et al., 2015)により，美しさ以外の 20 観点の評価が，美しさスコアを説明できるかを検討した
     - ✧ 385 人×40 俳句の 15,400 件の観測は，階層データおよび反復測定データの検定力に関する文献(Arend & Sch ¨afer, 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，独立変数は美しさ以外の 20 観点の評価，統制変数は学歴，変量効果は参加者，俳句，タスクの順序(評価ブロックが先か判別ブロックが先か)とした

＜生成者帰属＞

3. 「参加者が，人が生成した俳句と AI が生成した俳句(HOTL/HITL)を判別できるかを検証する」
   - ヒット率が偶然のレベル(0.5)と有意に異なるかどうかを *t* 検定で確認した

4. 「どのような要因，特に参加者の背景(教育や俳句の経験)や性格特性(ロボや AI に対する態度と関連するアニミズムや共感性)が，生成者の判別制度を説明するかを検証する」
   - 判別が正解かどうかを従属変数，個人の特性を独立変数，参加者と俳句を変量効果として，一般化線形混合モデル(ロジスティック解析)を分析した

4. Results

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

- Table 2 に俳句の各評価についての記述統計を示す
- 人条件，HOTL 条件，HITL 条件の美しさスコア(Table 2 の 1 行目)を比較した
  - 条件の主効果は $F(384, 2) = 212.41$, $p = .00$, $\eta 2 = 0.06$ で有意だった
    - ✧ 多重比較を行うと，HITL が最も高いスコアを示した
    - ✧ 人条件と ~~HITL~~ HOTL 条件では，スコアに差がなかった(Fig. 1 参照)
- また，Table 2 にはそのほかの 20 の観点の評価についての探索的分析の結果も示している
  - これらより，人が AI の出力に介入した場合(つまり HITL)，人が生成した俳句よりも高い評価を受けることが示唆された

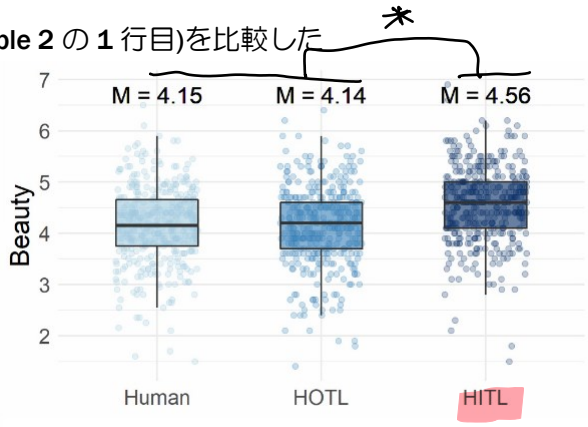

Fig. 1. Beauty scores in the human, HOTL, and HITL conditions.

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

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

- **20** 観点の美的評価，課題順序，俳句リスト，条件の要因が，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，最も説明力のある要因を決定した**(Table 3)**
  - 美的評価に関する **20** 観点では，生き生きとする**(*b* =0.01, *SE* =0.01, *t* =0.56, *p* =.57)**，心がさまよう**(*b* =0.01, *SE* =0.01, *t* =1.17, *p* =.24)** および体験との繋がり**(*b* =0.01, *SE* =0.01, *t* =1.31, *p* =.19)** を除いた全てが俳句の美しさに寄与した
  - 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，人と **HITL** の条件差**(*b* =0.18, *SE* =0.09, *t* =2.16, *p* =.03)** は俳句の美しさを説明した
- **4.1** 節で示したように，これらの結果から，人が介在する **AI** 生成俳句**(=HITL)**は，人が生成した俳句よりも美しさのスコアが高く，人が介在しない **AI** 生成俳句**(=HOTL)**のスコアは人が生成した俳句と同程度であることが示された
- 課題順序**(AI** が生成した俳句かどうかの事前知識**)**は，俳句の美しさの評価に影響しなかった
  - 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
  - 各条件の線形混合モデルの結果については，**Supplementary Table 1** を参照

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

- 生成者帰属(人が生成した俳句と **AI** が生成した俳句を判別する能力)については，人条件，**HOTL** 条件，**HITL** 条件のヒット率(正しく判断する確率)はそれぞれ**.55**，**.50**，**.43** だった(**Fig. 2**；各俳句とリストのヒット率は **Supplementary Table 2** も参照)
  - 人条件でのヒット率は $t(39) = 3.51$, $p = .001$ と有意に高いが，**HITL** 条件では $t(19) = -3.19$, $p = .005$ と有意に低い値である
  - **HOTL** 条件でのヒット率は偶然(.50)と差がなく，$t(19) = 0.03$, $p = .98$ であった
    - これらの結果から，参加者は **HOTL** の俳句について，人が生成しものと **AI** が生成したものの判別ができないことが示唆された
    - さらに，**HITL** 俳句の生成者は **AI** ではなく人であると信じていることを示唆した

- 美しいものは **AI** ではなく人が生み出すと信じている場合(=アルゴリズム嫌悪)，ヒット率と美しさスコアには関係があるかもしれない
  - そこで，各俳句の美しさスコアとヒット率の関係を探索的に検討した
    - その結果，人条件では $r(39) = .18$, $p = .26$ と正の相関を示したが有意ではなかった
    - 一方，**HOTL**，**HITL** 条件ではそれぞれ $r(19) = -.54$, $p = .01$, $r(19) = -.47$, $p = .04$ という有意な負の相関が見られた(**Fig. 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「美しい俳句ほど人が生み出したものである可能性が高い」と考えていることを反映している

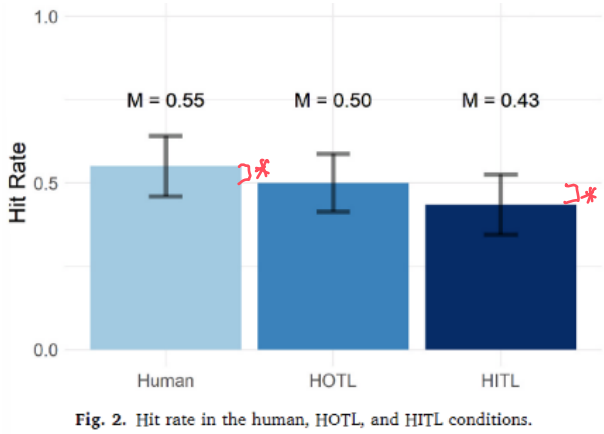

Fig. 2. Hit rate in the human, HOTL, and HITL conditions.

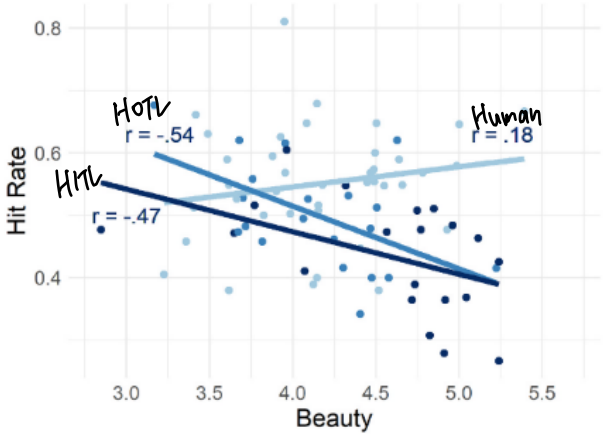

Fig. 3. The scatterplot between the beauty score and the hit rate.

4.4. the relationship between hit rate and personality traits

- 最後に，生成者の判別精度を説明する要因について検討した
  - その要因とは，学歴，俳句経験，アニミズムや共感性といった性格特性である
- Table 4 に示すように，アニミズムの擬人化度(*b* =.10, *SE* =0.03, *z* =2.81, *p* =.01)が高いほど，また俳句経験(*b* =0.07, *SE* =0.03, *z* =2.08, *p* =.04)が多いほど，正解率は高くなった
  - このことから，俳句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，俳句生成者の判別能力に影響を与えることが明らかとなった
  - 本研究で測定した全ての個人特性についての探索結果は，Supplementary Table 3 を参照

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

- 生成者判別の根拠(参加者がどのような手がかりに着目して俳句の生成者を判別したか)と，ヒット率の関係を探索的に検討した
  - 人が生成した俳句と判断する手がかりは「作品の深さ」(参加者の 72%)，AI が生成した俳句と判断する手がかりは「表現力」(参加者の 58%)が最も多かった
  - 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，β =0.21, *p* =.00)が人生成俳句のヒット率を説明できること，
  - 規則性(*b* =0.04, *SE* =0.02, *t* =2.24，β =0.12, *p* =.03)，意図性(*b* =0.09, *SE* =0.02, *t* =4.10，β =0.21, *p* =.00) およびその他(*b* = 0.15, *SE* =0.04, *t* =3.39，β = 0.17, *p* =.00)により AI 生成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，明確なルールを持ち，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，人と AI が生成した俳句を様々な角度から評価した
  - その結果，人が選んだ AI の俳句は，人が生成した俳句よりも美しさスコアが高いことを示した
  - ランダムに選ばれた AI の俳句は，人が作った俳句と同程度の美しさであることを示した
    - これらの結果は，人と AI の共同作業がより良い創造性につながること，また，少なくとも俳句の生成においては，AI の生成力が創造的な分野において人と同等であることを示唆している
  - また，俳句の美しさを説明する要因として，哲学的な美的と心理学的な美的感覚の両方があることも明らかになった
  - 生成者の判別については，参加者は人が生成した俳句と AI が生成した俳句を判別することができなかった
    - この結果と上述の美しさスコアは，AI が人と同等の品質の作品を生み出せ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，人が生成した俳句と AI が生成した俳句の判別能力に寄与していると考えられる

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

- 本研究の結果は，人が生成した俳句と，人が介在しない AI が生成した俳句(HOTL)の美しさスコアが同程度であったことを示している
  - この結果は，人が生成した詩が最も好まれた先行研究(Köbis & Mossink, 2021)と矛盾する
    - 一つの説明は，詩のスタイルの違いによるものと考えられる
      - 俳句は世界的に見ても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，このような状況でも，AI アートのクオリティは十分に確保されていると考えられる
      - AI の訓練や訓練用の材料が，過去の研究で使用されたものよりも優れていたと思われることも特筆すべき点である
        - 俳句は，5-7-5 音節の形式や季語を入れるなどの明確なルールがあることが特徴

で，トレーニングが容易だったのかもしれない

- ✧ 他の説明は，方法論の違いによるものと考えられる
  - 先行研究では，AI が生成した詩の最初の 2 行を人が作った詩と同じにし，両方の詩を並べ，参加者にどちらが好きかを尋ねている(つまり，2 択の強制選択である)(K ¨obis & Mossink, 2021)
  - 二者択一の強制選択は，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
    - ➢ 仮にそうであっても，人が生成した作品と AI が生成した作品の差は微妙である
  - 本研究では，方法論が少し異なることで，人が生成した作品にあった優位性が，非専門家にとって失われることを示した

5.2. Factors explaining beauty

- 本研究は，人が生成した俳句と AI が生成した俳句の美しさやヒット率を説明するいくつかの要因を明らかにした
  - ➢ 先行研究の限界の一つは，好感度や美しさといった限定的な項目で人の詩と AI の詩を比較していることである
    - ✧ しかし，本研究ではこの制約を克服した
      - 例えば，人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，ポジティブな感情，畏怖や懐かしさといった高次の感情(Hitsu-war & Nomura, 2022b; 2022c)が，AI 俳句の美しさについても説明要因になることが明らかになった
      - また，美しさの実証研究が始まる以前から，哲学者や思想家が美しさを説明すると考えられてきた哲学的な要因(Brielmann et al., 2021)も，詩の美しさに関係することが再び明らかになった

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

- 人が生成した詩と AI が生成した詩が判別されなかったことは，これまでの知見(Hopkins & Kiela, 2017; Lau et al., 2018; Schwartz, 2015)と一致する
  - ➢ 他の詩のタイプと同様に，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)
  - ➢ さらに，一般的な詩よりもはるかに短く，個人の想像に依存するという俳句特有の性質が，人が介在しない AI 生成の俳句においても，その判別を阻害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負の関係を考慮すると，人はアルゴリズム嫌悪を持っている可能性がある(Burton et al., 2020)
  - ➢ アルゴリズム嫌悪は，人間と AI の詩を比較するという研究文脈において重要な概念である (K ¨obis & Mossink, 2021)
  - ➢ ここでは，AI 俳句を見下した考え方があるのかもしれない
  - ➢ 本研究は，アルゴリズム嫌悪と生成者判別の関係性を示唆する初めての研究であり，この結果は，AI アートに対する現代人の意識を反映していると考えられる

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

- 擬人化傾向や共感といった性格特性要因は，AI やロボットとの相互作用の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，それらが俳句作品の生成者判別においても意味を持つことを示した
    - この結果から，擬人化傾向の高い人は，美しい俳句を衝動的に人が生成したものと判断せず，結果としてヒット率が高くなることが推測される
    - さらに，俳句に使われている言葉が一致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が生成した俳句についてヒット率が高いことが示された
    - これに対し，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，意味の一貫性がなく，理解しにくいものがあったため，規則性に注目することが，人が生成した俳句を判別するためのツールとして機能した
      - また，AI アートの鑑賞では規則性が認められることが多く(cf. Chamberlain et al., 2018)，意図性の有無は AI アート研究の大きなテーマの一つである(Chamberlain et al., 2018; McCormack et al., 2019)
        - これらの要因と AI アート検出の関係性は指摘されているものの，実証されていない
        - 本研究は，この点をデータとして示した初めての研究である

5.5. Limitations and future research

- 本研究には，いくつかの制約があった
  - そのーつは，HITL 条件における AI 俳句の選定方法である
    - 本研究では，選定は専門家ではない 3 名が行った
      - 参加者の多くが非専門家であったため，専門家が選択した人生成の俳句よりも HITL 俳句の方が理解が容易であったかもしれない
    - 今後は，非専門家が選択した人生成の俳句や，専門家が介入した HITL 俳句を用いた実験が行われる可能性もある
    - それと同時に，俳句初心者や俳句の専門家が，生成された俳句の美しさを評価したり，生成者を判別したりする検討も可能である
      - とはいえ，少なくとも素人の視点に立つと，本研究は，人と AI の共同作業が素晴らしい作品を生み出す可能性を持っていることを示している
- なお，本研究では，日本人の参加者に対して，日本語の俳句のみを提示した
  - 海外では俳句生成の AI モデルが洗練されている(Hreˇskov´a & Machov´a, 2018; Wong, Chun, Li, Chen, & Xu, 2008)
  - さらに，俳句の東西文化比較では，美しさ評価に有意差はなかった(Hitsuwari & Nomura, 2022c)
    - したがって，本研究は他の文化にも適用できる可能性がある

5.6. Conclusions

- 本研究では，AI 生成の俳句と人生成の俳句の美的評価と生成者の判別を検討した
  - 美しさスコアは，人が生成した俳句とランダムに選択した AI 生成の俳句において同程度だった
    - しかし，AI が生成した俳句でも，人が選択したものであれば，美しさスコアが最も高かった
  - さらに，人が生成した俳句と AI が生成し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
  - また，AI の俳句では，美しさスコアと判別制度の間に負の相関があることが分かった
    - これらの結果を総合すると，(情報が少ない)俳句において，AI アートの品質は人と遜色ないレベルに達しており，人と AI の共同作業により，より創造的なアートワークが生み出されることが示唆される(Booten & Gero, 2021)
    - この知見は，AI が人の生成を支援することで創造性が促進されることを示唆しており，アートなどの創造性に関連する領域に影響を与える可能性がある