

生成 AI による心理学研究用顔画像データベースの作成

研究代表者 井 上 和 哉 東京都立大学人文社会学部・准教授

1 本研究の背景

本研究の目的は、画像生成 AI サービスを用いて、心理学や情報学を含む多様な学術研究で利用可能な日本人顔画像データベースを作成し、その妥当性を実証的に検討することである。顔画像は、注意、記憶、魅力、表情認知、社会的印象形成などの心理学研究において広く用いられる基本的な刺激である。同時に、顔画像は、バナー広告、ソーシャルネットワーキングサービスのプロフィール画像、オンライン上の人物印象形成など、情報通信に関わる人文社会科学的研究でも頻繁に用いられる。したがって、研究者が自由に利用でき、かつ表情や印象に関する妥当化情報が付与された顔画像データベースを整備することは、心理学研究だけでなく、多様な分野の学術的基盤整備としても重要である。

顔画像を用いた研究では、研究目的に応じて多数の人物画像を必要とする。顔認知には人種や文化、生活環境の影響が認められるため、日本人参加者を対象とする研究では、日本人顔画像を刺激として用いることが望ましい。欧米では KDEF, NimStim, Radboud Faces Database などの顔表情画像データベースが整備され、表情分類、強度評定、魅力度評定などに基づく妥当化が行われてきた (Goeleven et al., 2008; Tottenham et al., 2009; Langner et al., 2010)。しかし、日本人顔画像を主とする研究用データベースは十分に多いとはいえない。FIND, ATR 顔表情データベース, AIST 顔表情データベース, KRC 顔表情データベースなど、日本人顔画像を含むデータベースは存在するものの、公開状況、モデル数、表情カテゴリ、付随するバリデーション情報の点で制約が残る (蒲池, 2001; 渡邊他, 2007; Fujimura & Umemura, 2018; Ueda et al., 2019)。その結果、日本の研究者は、研究室ごとに独自の未公開刺激を作成したり、海外データベースを代替的に使用したりせざるを得ない場合があり、研究結果の再現性や比較可能性を低下させる要因となっている。

したがって、顔画像を使用する学術研究を推進するには、十分なモデル数及びバリデーション情報を含み、権利上問題なく公開可能な顔画像データベースが必要である。しかし、実在人物を撮影して顔画像データベースを作成するには、モデル募集、撮影環境の統制、表情演技の指導、肖像権や公開許諾の取得など、多くのコストを要する。とくに顔画像を広く公開する場合、モデル本人が自分の顔が不特定多数の研究者に利用されることを許諾する必要がある、データベースの規模拡大を妨げる大きな要因となる。

近年発展している画像生成 AI は、この問題を緩和する可能性をもつ。AI によって生成された顔画像は実在顔画像と区別しにくく、場合によっては実在顔よりも本物らしく知覚されることが報告されている (Nightingale & Farid, 2022; Miller et al., 2023; Bray et al., 2023)。また、深層学習により合成されたアジア人顔画像データセットを作成し、評定実験により妥当性を検討する試みも出てきている (Han et al., 2023)。このような知見を踏まえると、生成 AI を用いて研究用顔画像刺激を作成し、バリデーションデータとともに整備することは、今後の顔研究にとって有効な方法となりうる。

一方で、生成 AI によって作成された顔画像が、心理学実験の刺激として直ちに妥当であるとは限らない。例えば表情画像を作成し利用するためには、その画像が意図した表情カテゴリとして分類されるか、またその表情がどの程度強く知覚されるかを確認する必要がある。既存の表情画像データベースでも、表情分類課題や強度評定課題を用いて刺激の妥当性が検証されている (Goeleven et al., 2008; Tottenham et al., 2009; Ueda et al., 2019)。そこで本研究では、生成 AI によって作成した日本人顔表情画像について、表情分類課題と表情強度評定課題を実施し、刺激としての妥当性を検討した。さらに、分類正答率と強度評定値の関連を刺激単位で検討し、生成 AI 表情画像を実験刺激として選別するための基礎情報を得ることを目指した。

2 方法

2-1 参加者

表情分類課題および表情強度評定課題には、同一の 31 名が参加した。内訳は女性 15 名、男性 16 名であり、平均年齢は 21.77 歳 (SD = 3.96, 範囲 : 18- 35 歳) であった。

なお、両課題は同一参加者に対して実施されたが、いずれの課題においても注意確認試行を設け、課題に適切に取り組んでいるかを確認した。注意確認試行の詳細は各課題の結果で述べる。

2-2 刺激

本研究では、生成 AI によって作成した日本人顔画像を刺激として用いた。顔画像の作成には、Adobe 社の画像生成 AI である Firefly を用いた。具体的には、男性・女性の各モデルにつき、怒り、嫌悪、恐怖、幸福、悲しみ、驚きからなる基本 6 表情 (Ekman & Friesen, 1971) と中立表情の計 7 表情を作成した。それぞれの表情を作成する際には、基本的には各感情語をプロンプトに含む形で作成した。背景は白で、服装は黒に統一し、作成した顔は全て正面を向いていた。2 名の研究補助員が独立して顔画像を作成し、顔や表情の自然さ、表情の適切さ、人物としての一貫性、背景の統一性などに関して、補助員相互の確認及び研究代表者の最終確認を行った上で、画像の予備評価に使用する顔画像を選定した。画像の予備評価フェイズでは、一人の参加者につき、男女 30 モデル×7 表情の顔を 1 枚ずつ提示し、表情分類を求めた。表情分類の成績が著しく低かったモデルの顔画像や、再度のチェックにより不適切であるとみなされた顔画像はデータベースの作成から除外された。なお、煩雑になるため、予備評価フェイズの細かい結果は省略する。

以上の手続きを繰り返し何度も行い、最終的な刺激セットは 100 名分の人物画像から構成され、その内訳は女性顔 50 名、男性顔 50 名であった。各人物について、怒り、嫌悪、恐怖、幸福、中立、悲しみ、驚きの 7 表情、合計 700 枚の表情画像を整備した。表情分類課題では 7 表情すべてを用いた。表情強度評定課題では、強度評定の対象となる怒り、嫌悪、恐怖、幸福、悲しみ、驚きの 6 表情を用い、中立表情は含めなかった。

本研究で作成した刺激数は、表情分類課題に用いた正面顔・7 表情の刺激セットとして 700 枚であり、既存の日本人顔表情データベースのバリデーション実験で用いられる刺激数と比較しても大規模である。たとえば KRC 顔表情データベースでは、正面・直視条件の 7 表情画像として 74 名分、計 518 画像がバリデーション実験に用いられている (Ueda et al., 2019)。本研究では、生成 AI を用いることで、実写撮影に伴う肖像権や撮影コストの問題を回避しつつ、100 名分の 7 表情画像を作成し、実験的に評価することを可能にした。

2-3 表情分類課題

表情分類課題では、参加者に顔画像を 1 枚ずつ提示し、その顔がどの表情を表しているかを選択させた。選択肢は、怒り、嫌悪、恐怖、幸福、中立、悲しみ、驚きの 7 種類であった。各参加者は 100 名分の 7 表情画像を判断したため、本試行数は 1 人あたり 700 試行であった。刺激の提示順は参加者ごとにランダム化した。参加者の疲労を考慮し、50 試行ごとに休憩を行った。

課題中には注意確認試行を挿入した。注意確認試行では、顔画像の代わりに黒い四角形を提示し、スペースキーを押すよう求めた。この試行は、参加者が画像を見ずに反応を続けていないかを確認するためのものであった。

2-4 表情強度評定課題

表情強度評定課題では、参加者に各顔画像の表情の強さを評定させた。評定対象は怒り、嫌悪、恐怖、幸福、悲しみ、驚きの 6 表情であり、中立表情は評定対象に含めなかった。課題は表情ごとのブロックで構成され、参加者は各ブロックにおいて、当該表情がどの程度強く表れているかを 7 件法で評定した。評定尺度は、1 を「非常に弱い」、4 を「中程度」、7 を「非常に強い」とした。

各表情ブロックでは 100 枚の顔画像が提示され、6 表情で合計 600 試行が行われた。ブロック順および各ブロック内の刺激提示順は参加者ごとにランダム化した。表情分類課題と同様に、課題中には黒い四角形と「スペースキーを押してください」という指示を提示する注意確認試行を挿入した。表情分類課題と同様に、参加者の疲労を考慮し、50 試行ごとに休憩を行った。

2-5 分析方針

表情分類課題では、各参加者について表情ごとの分類正答率を算出し、その後、参加者間で平均した。分類成績が 7 択判断のチャンスレベルである 14.29%を上回るかを確認するため、参加者ごとの表情別正答率を用いて表情ごとに 1 標本 t 検定を行った。また、混同行列を作成し、どの表情間で誤分類が生じやすいかを確認した。さらに、KRC 顔表情データベースの分類成績を固定された基準値として扱い、本研究の分類正答率と比較した。

表情強度評定課題では、各参加者について表情ごとの平均評定値を算出し、その後、参加者間で平均した。尺度の理論的中点である 4 との比較には、参加者ごとの表情別平均評定値を用いた 1 標本 t 検定を行った。

また、KRC 顔表情データベースの強度評定値を固定された基準値として扱い、本研究の評定値と比較した。さらに、刺激単位で分類正答率と平均強度評定値を結合し、表情ごとに Pearson の積率相関係数を算出した。これにより、同一表情カテゴリ内で、表情が強く知覚される刺激ほど、意図した表情として正しく分類されやすいかを検討した。

3 結果

3-1 表情分類課題の結果

表情分類課題では、31 名全員が 700 試行を完遂した。注意確認試行は各参加者ごとに 14 試行であり、すべての試行で正答が得られた（正答率 100%）。したがって、注意確認試行の成績に基づいて除外した参加者はいなかった。

全試行を通した分類正答率は 76.6%であった。参加者ごとに各表情の分類正答率を算出し、その後、参加者間で平均した結果を図 1 に示す。分類正答率は表情カテゴリによって大きく異なっていた。幸福は 98.7%，中立は 95.4%，驚きは 91.2%と高い分類正答率を示した。怒りは 82.1%，悲しみは 81.2%であり、これらも比較的高い水準であった。一方、嫌悪は 52.8%，恐怖は 34.8%であり、他の表情に比べて低かった。

各表情の分類正答率がチャンスレベル 14.29%を上回るかを検討したところ、すべての表情でチャンスレベルを有意に上回った（Holm 補正後 $ps < .001$ ）。したがって、嫌悪や恐怖の正答率は相対的に低かったものの、いずれの表情もチャンスレベルを超えて意図した表情として分類されていた。

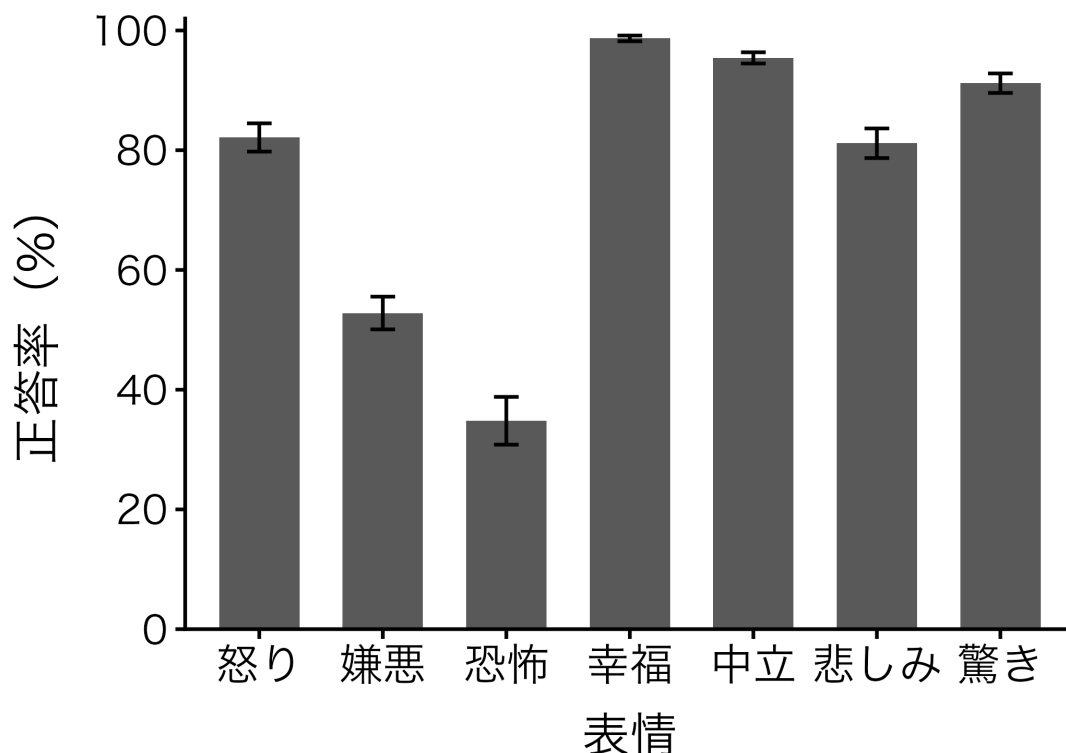


図 1 表情分類課題における表情別分類正答率。エラーバーは標準誤差を示す。

各表情がどの表情として誤分類されたかを検討するために、混同行列を作成し、表 1 に示す。混同行列の対角成分は、各表情が意図したカテゴリとして分類された割合を示す。幸福、中立、驚きでは対角成分が高く、誤分類は比較的少なかった。一方、嫌悪では、嫌悪として分類された割合に加えて、怒りや悲しみとして分類される反応も多かった。また、恐怖では、恐怖として分類された割合よりも、驚きとして分類された割合が高く、恐怖と驚きの混同が顕著であった。

表情分類課題の混同行列

行：作成時に意図した表情，列：参加者の回答表情。セルは割合（％）を示す。

正解表情	参加者の回答表情						
	怒り	嫌悪	恐怖	幸福	中立	悲しみ	驚き
怒り	82.1	11.7	2.4	0.1	0.3	1.5	1.8
嫌悪	24.1	52.8	2.3	0.0	3.1	16.1	1.5
恐怖	8.0	7.3	34.8	0.1	0.7	3.4	45.8
幸福	0.1	0.1	0.0	98.7	0.9	0.1	0.1
中立	1.8	0.5	0.1	0.9	95.4	0.4	0.9
悲しみ	1.4	8.0	0.4	0.1	8.7	81.2	0.2
驚き	1.4	0.1	0.9	3.5	2.7	0.1	91.2

各行の割合は、その表情として作成された刺激に対する回答分布を示す。対角成分は正答分類率である。

表 1 表情分類課題における混同行列。行は作成時に意図した表情，列は参加者が選択した表情を示す。各セルの値は反応割合（％）である。

KRC 顔表情データベースの分類正答率と比較した結果を表 2 に示す。KRC の正分類率を基準値として扱い、本研究の参加者ごとの表情別分類正答率を用いて 1 標本 t 検定を行った。その結果、怒り、嫌悪、幸福、中立、悲しみでは、本研究の分類正答率が KRC の値を有意に上回った。恐怖と驚きでは有意差は認められなかった。とくに恐怖については、本研究の正答率が KRC とほぼ同程度であり、恐怖表情の分類困難性が本研究に固有の問題ではないことを示唆している。

KRCデータベースの分類正答率との比較

KRCの表情別正答率を固定された基準値として扱った1標本t検定。

表情	参加者数	KRC正答率	本研究正答率	差	t	df	p	Holm補正p	dz	方向
怒り	31	38	82.1	44.1	18.70	30	<.001	<.001	3.36	本研究 > KRC
嫌悪	31	41	52.8	11.8	4.32	30	<.001	<.001	0.78	本研究 > KRC
恐怖	31	34	34.8	0.8	0.20	30	.841	.939	0.04	本研究 > KRC
幸福	31	86	98.7	12.7	26.63	30	<.001	<.001	4.78	本研究 > KRC
中立	31	93	95.4	2.4	2.58	30	.015	.045	0.46	本研究 > KRC
悲しみ	31	53	81.2	28.2	11.39	30	<.001	<.001	2.05	本研究 > KRC
驚き	31	90	91.2	1.2	0.73	30	.469	.939	0.13	本研究 > KRC

KRC側のサンプリング誤差は考慮していないため、既存データベースを基準とした探索的・記述的比較として解釈する。

表 2 KRC 顔表情データベースとの分類正答率の比較。KRC 値を固定された基準値として扱った 1 標本 t 検定の結果を示す。

3-2 刺激単位でみた分類成績のばらつき

次に、各刺激について分類正答率を算出し、表情ごとに刺激単位でのばらつきを確認した。図 2 に示すように、同じ表情カテゴリ内でも刺激間に大きな差がみられた。幸福や中立では多くの刺激が高い分類正答率を示したが、嫌悪や恐怖では刺激ごとのばらつきが大きかった。恐怖では、ほとんど正しく分類されない刺激から、比較的高い正答率を示す刺激までが混在していた。

この結果は、表情カテゴリ単位の平均値だけでは、個々の刺激の妥当性を十分に判断できないことを示している。生成 AI によって作成された表情画像を研究刺激として用いる場合には、表情カテゴリ名だけに基づいて刺激を選ぶのではなく、刺激単位の分類成績を確認する必要がある。

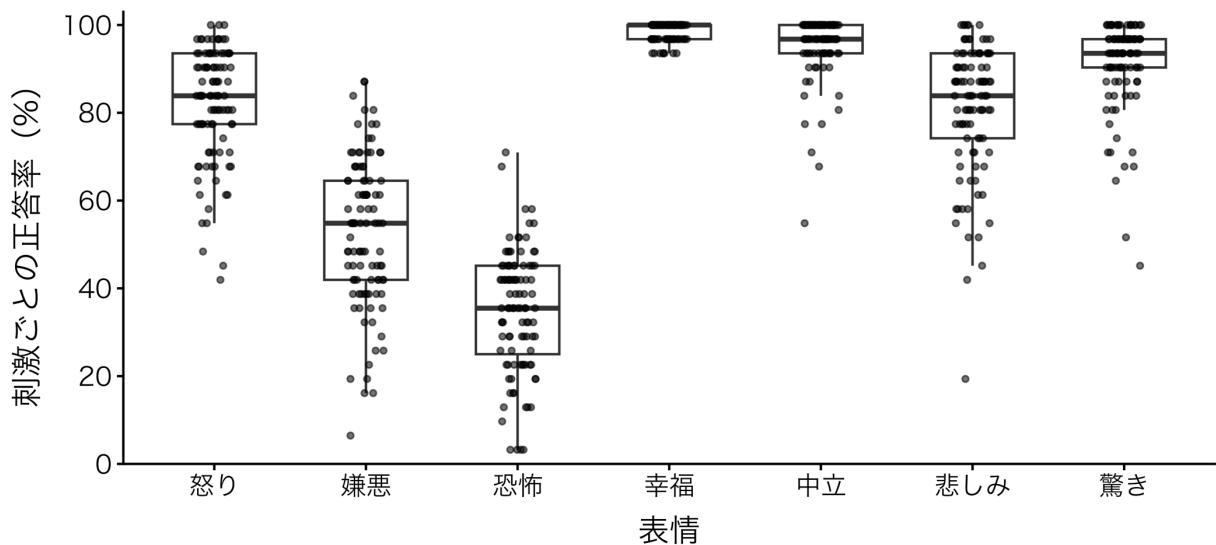


図 2 刺激単位でみた分類正答率の分布。各点は 1 刺激の分類正答率を示す。

3-3 表情強度評定課題の結果

表情強度評定課題では、31 名全員が 600 試行を完遂した。注意確認試行は参加者ごとに 12 試行であり、すべて正答であった（正答率 100%）。したがって、強度評定課題においても注意確認試行の成績に基づく除外者はいなかった。

本研究の表情強度評定値と KRC 顔表情データベースの強度評定値を図 3 に示す。図中の丸印は本研究の平均評定値、×印は KRC データベースの強度評定値を示す。本研究では、怒りが 4.48、恐怖が 4.29、幸福が 4.16、驚きが 4.11、嫌悪が 4.07 であり、多くの表情は 7 件法尺度の midpoint 4 付近からやや上の値を示した。一方、悲しみは 3.61 であり、尺度 midpoint を下回った。

7 件法尺度の midpoint 4 との比較結果を表 3 に示す。Holm 補正後、怒りと恐怖は尺度 midpoint を有意に上回った。一方、悲しみは尺度 midpoint を有意に下回った。嫌悪、幸福、驚きは平均値としては 4 付近に位置していたが、Holm 補正後には尺度 midpoint との差は有意ではなかった。したがって、本研究で作成した表情画像は全体として中程度前後の表情強度を有していたが、表情カテゴリによって強度の明確さには差があった。

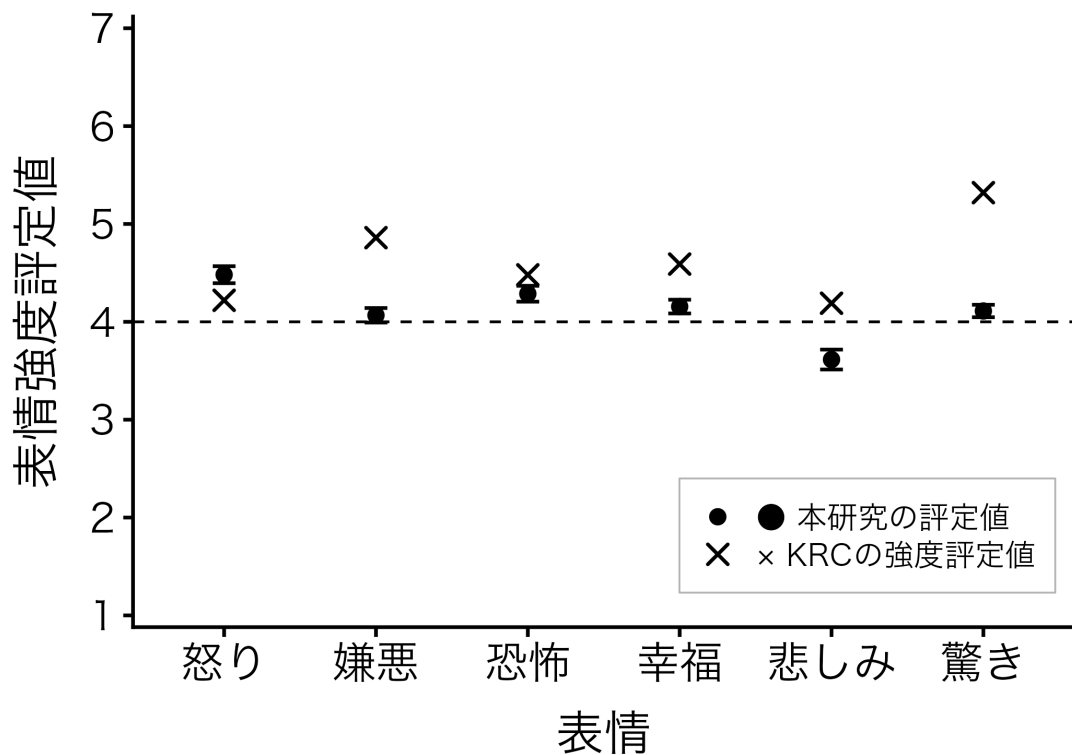


図3 表情強度評定値と KRC 顔表情データベースとの比較。丸印は本研究の平均評定値，×印は KRC の強度評定値を示す。エラーバーは本研究における標準誤差である。

尺度中点4との比較											
参加者ごとの表情別平均評定値を用いた1標本t検定。											
表情	参加者数	中点	平均強度	差	t	df	p	Holm補正p	95%CI差	dz	方向
怒り	31	4	4.48	0.48	5.55	30	< .001	< .001	0.3-0.66	1.00	中点より高い
嫌悪	31	4	4.07	0.07	0.90	30	.375	.375	-0.08-0.22	0.16	中点より高い
恐怖	31	4	4.29	0.29	3.54	30	.001	.005	0.12-0.45	0.64	中点より高い
幸福	31	4	4.16	0.16	2.22	30	.034	.103	0.01-0.3	0.40	中点より高い
悲しみ	31	4	3.61	-0.39	-3.79	30	< .001	.003	-0.59-0.18	-0.68	中点より低い
驚き	31	4	4.11	0.11	1.74	30	.092	.185	-0.02-0.24	0.31	中点より高い

p値は両側検定。Holm補正pは6表情に対する多重比較補正後の値。

表3 表情強度評定値と尺度中点4との比較。参加者ごとの表情別平均評定値を用いた1標本t検定の結果を示す。

KRC との比較では、怒りのみ本研究の評定値が KRC 値を上回り、嫌悪、恐怖、幸福、悲しみ、驚きでは本研究の評定値が KRC 値を下回った。とくに驚きでは、本研究の評定値が KRC よりも大きく低く、生成 AI によって作成された驚き表情は、KRC の驚き表情ほど強くは知覚されなかった可能性がある。ただし、KRC では各画像に対して6表情すべての強度を評定しており、その表情がほとんど含まれない画像も比較対象に含まれていた。一方、本研究では表情ごとのブロック内で当該表情画像のみを評定した。したがって、KRC との比

較は、厳密な直接比較ではない点には留意が必要である。

3-4 刺激単位でみた表情強度のばらつき

各刺激について 31 名の平均評定値を算出し、表情ごとに表情強度の分布を確認した（図 4）。表情強度についても、分類正答率と同様に、同じ表情カテゴリ内で大きなばらつきが認められた。怒り、嫌悪、恐怖、悲しみ、驚きでは、刺激ごとの平均評定値の分布が広く、弱く知覚される刺激と強く知覚される刺激が混在していた。幸福は相対的にばらつきが小さいものの、それでも刺激単位で一定の差がみられた。

この結果は、生成 AI 画像の表情強度が、表情カテゴリ内で均一ではないことを示している。したがって、表情刺激を選定する際には、表情カテゴリ名だけでなく、各刺激の強度評定値を参照することが必要である。特定の研究目的に応じて、強い表情刺激を選ぶ場合もあれば、中程度の強度をもつ刺激を選ぶ場合もあるため、本研究で得られた刺激単位の強度評定データは、今後の刺激選別に利用できる基礎資料として有用である。

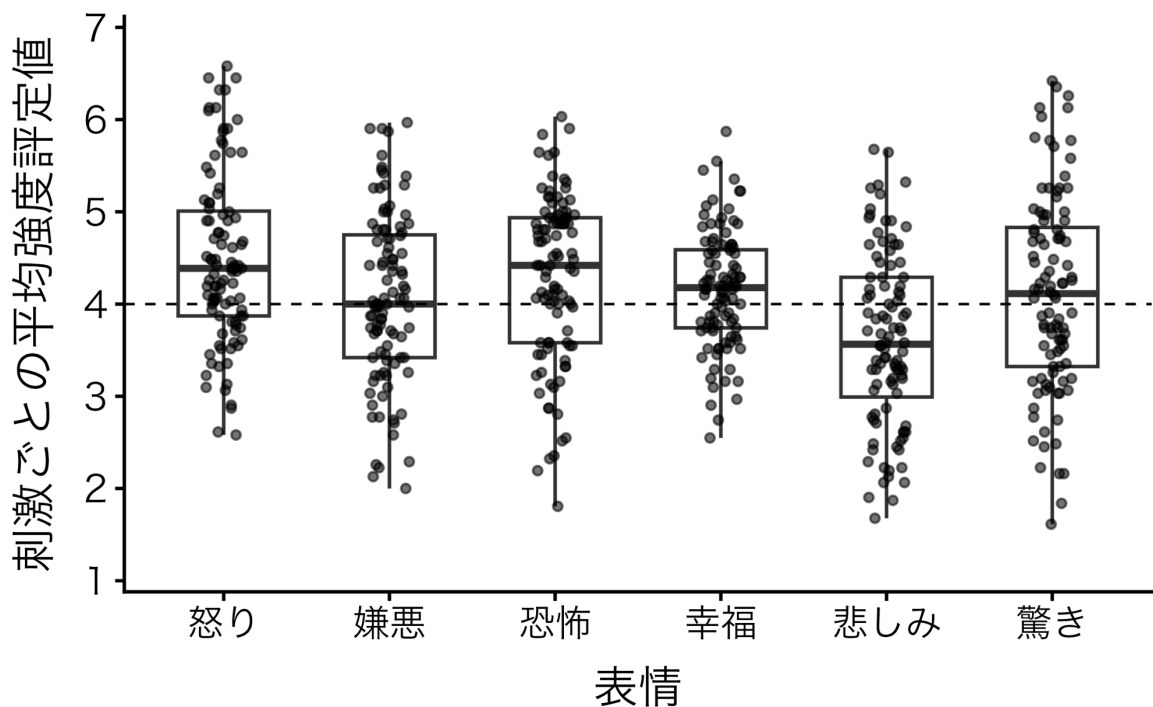


図 4 刺激単位でみた表情強度評定値の分布。各点は 1 刺激の平均評定値を示す。破線は尺度中点 4 を示す。

3-5 分類正答率と表情強度評定値の関連

表情分類課題における刺激ごとの分類正答率と、表情強度評定課題における刺激ごとの平均強度評定値との関連を検討した。分析単位は刺激であり、強度評定課題に含まれない中立表情を除外したうえで、怒り、嫌悪、恐怖、幸福、悲しみ、驚きの 6 表情、合計 600 刺激を対象とした。各刺激について、分類課題では意図した表情として正しく分類された割合を算出し、強度評定課題では当該表情の平均評定値を算出した。

表情ごとの散布図を図 5 に、Pearson の相関係数を表 4 に示す。怒り、幸福、悲しみ、驚きでは、刺激ごとの強度評定値と分類正答率の間に有意な正の相関が認められた。とくに怒りと悲しみでは比較的強い正の相関が示され、表情が強く知覚される刺激ほど、意図した表情として正しく分類されやすかった。一方、嫌悪と恐怖では有意な相関は認められなかった。

この結果は、表情強度評定値と分類正答率の対応が表情カテゴリによって異なることを示している。怒り、悲しみ、幸福、驚きなどの分類正答率が高い表情では、強度評定値が刺激選別の有用な指標になりうる。一方、嫌悪や恐怖では、表情の強さだけでなく、他表情との類似性や混同されやすさが分類成績に影響している可能性がある。したがって、刺激選別においては、強度評定値だけでなく、分類正答率や混同パターンも併せて確認することが重要である。

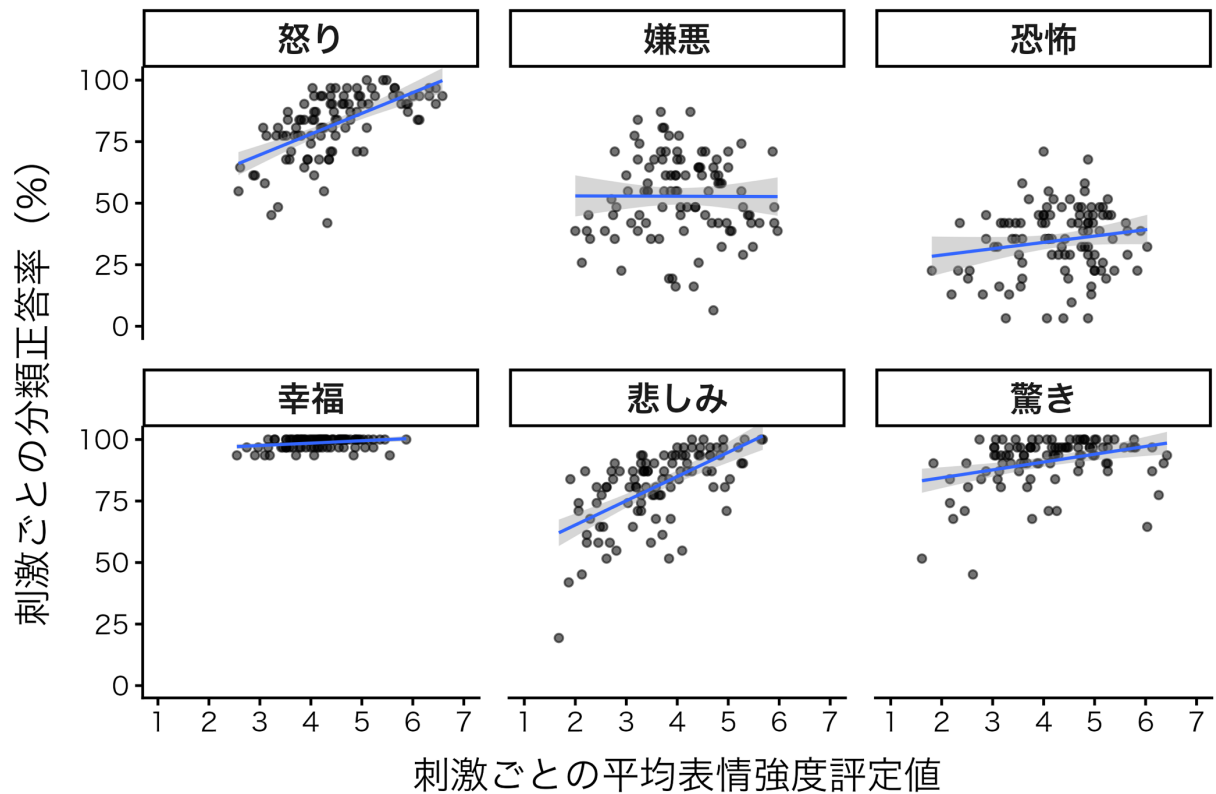


図5 表情ごとの分類正答率と表情強度評定値の関連。各点は1刺激を示す。

刺激単位の分類正答率と表情強度評定値の相関：表情別				
表情	刺激数	Pearson_r	p値	Holm補正p
怒り	100	0.608	< .001	< .001
嫌悪	100	-0.004	.971	.971
恐怖	100	0.167	.097	.194
幸福	100	0.298	.003	.008
悲しみ	100	0.620	< .001	< .001
驚き	100	0.335	< .001	.003

相関係数は Pearson の積率相関係数である。Holm補正pは、6表情に対する多重比較補正後の値である。

表4 刺激単位の分類正答率と表情強度評定値の Pearson 相関。Holm 補正 p は6 表情に対する多重比較補正後の値である。

3-6 高妥当性刺激と低妥当性刺激の例示

分類正答率と表情強度評定値を組み合わせ、刺激ごとの妥当性スコアを算出した。具体的には、表情カテゴリごとに分類正答率と強度評定値を標準化し、それらの合計を妥当性スコアとした。これにより、もと

もと分類されやすい表情や強く評定されやすい表情の影響を一定程度調整したうえで、同一表情カテゴリ内における相対的に妥当性の高い刺激と低い刺激を抽出した。

表情ごとの高妥当性刺激の例を図6に、低妥当性刺激の例を図7に示す。高妥当性刺激では、分類正答率と表情強度評定値がともに高く、意図した表情として比較的明確に知覚されている刺激が選ばれた。一方、低妥当性刺激では、表情が弱く知覚されたり、他表情として分類されやすかったりする刺激が含まれていた。成功例と低妥当性例を並べて示すことで、生成AIによって作成された表情画像には有用な刺激が多数含まれる一方、無選別のまま使用するには注意が必要であることが視覚的にも確認できる。



図6 表情ごとの高妥当性刺激の例。分類正答率と表情強度評定値に基づく妥当性スコアが高い刺激を示す。



図7 表情ごとの低妥当性刺激の例。分類正答率または表情強度評定値が相対的に低い刺激を示す。

4 考察

本研究では、生成 AI によって作成した日本人顔表情画像について、表情分類課題と表情強度評定課題を実施し、心理学研究用刺激としての妥当性を検討した。その結果、生成 AI によって作成された表情画像は、全体としてチャンスレベルを大きく上回る分類成績を示し、多くの表情において中程度前後の強度として知覚された。このことは、生成 AI を用いて、日本人顔の表情画像刺激を比較的大規模に作成し、実験刺激として利用できることを示している。

ただし、妥当性は表情カテゴリによって様ではなかった。幸福、中立、驚き、怒り、悲しみは比較的高い分類正答率を示した一方で、嫌悪と恐怖では正答率が相対的に低かった。とくに恐怖では、驚きへの誤分類が多く認められた。恐怖と驚きの混同は、本研究に固有の現象ではなく、先行研究でも報告されている。恐怖と驚きは、眉の上昇、開眼、口の開きなどの顔面筋活動を共有しており、恐怖に特有の手がかりが十分に知覚・利用されない場合、両表情の弁別が困難になると考えられている (Roy-Charland et al., 2014)。したがって、本研究で観察された恐怖から驚きへの誤分類は、生成 AI 画像の単なる失敗というよりも、恐怖表情と驚き表情の視覚的類似性を反映したものと考えられる。

表情強度評定課題では、怒りと恐怖が尺度中点を上回った一方、悲しみは尺度中点を下回った。この結果は、生成 AI によって作成された悲しみ表情が、相対的に弱く知覚される傾向を示している。悲しみは分類課題では一定の正答率を示したが、強度評定では必ずしも強い表情とは評価されなかった。これは、悲しみ表情が比較的微妙な顔面手がかりに基づいて知覚されるためである可能性がある。悲しみ表情には、眉の内側の上昇、眉の寄り、口角の下制などの特徴が関与するとされるが、これらの手がかりが十分に強調されていない場合、観察者はその表情を悲しみとして分類できても、表情強度としては低く評定する可能性がある。

したがって、悲しみ表情については、分類正答率だけではなく、強度評定値に基づく刺激選別や追加生成が必要である。

分類正答率と強度評定値の関連を刺激単位で検討した結果、怒り、悲しみ、幸福、驚きでは、表情が強く知覚される刺激ほど意図した表情として分類されやすかった。一方、嫌悪と恐怖では、強度評定値と分類正答率の明確な関連は認められなかった。このことは、表情の強さが高ければ必ずしも分類妥当性が高いとは限らないことを示している。嫌悪や恐怖では、他表情との混同や表情カテゴリとしての典型性が分類成績に影響している可能性がある。したがって、刺激選別においては、強度評定値、分類正答率、混同行列を組み合わせて判断することが重要である。

刺激単位の分析からは、同じ表情カテゴリ内であっても、妥当性に大きな差があることが示された。これは、生成 AI が感情語に対応した画像を生成できる一方で、個々の画像における表情の典型性や強度、他表情との弁別性までは安定して保証しないことを意味する。したがって、生成 AI による表情画像を研究刺激として使用する場合には、表情カテゴリ名に基づいて一括して採用するのではなく、分類正答率や強度評定値を参照し、刺激単位で利用目的に適した画像を選ぶことが重要である。本研究で示した高妥当性刺激と低妥当性刺激の例は、このような刺激単位の評価が、実際の刺激選別に有用であることを視覚的にも示している。

5 本研究から得られた成果と今後の課題

本研究の成果は、生成 AI によって作成された日本人顔表情画像が、心理学実験刺激として一定の妥当性を有することを示した点にある。100 名分の 7 表情、合計 700 画像を作成し、表情分類課題と表情強度評定課題によって妥当性を評価した。その結果、多くの表情でチャンスレベルを上回る分類成績が得られ、参加者は生成 AI 画像から意図された表情を一定程度読み取ることができた。また、強度評定課題からは、生成 AI 表情画像が全体として中程度前後の表情強度をもつことが示された。

同時に、本研究は、生成 AI 表情画像の妥当性が一様ではないことも明らかにした。嫌悪と恐怖では分類正答率が低く、悲しみでは強度評定値が低かった。また、同じ表情カテゴリ内でも刺激間のばらつきが大きく、強く知覚される刺激と弱く知覚される刺激が混在していた。これらの結果は、生成 AI 画像を研究刺激として用いる際に、表情カテゴリ名だけに基づいて刺激を選ぶのではなく、分類正答率、強度評定値、混同パターンに基づいて刺激を選別する必要があることを示している。

今後の課題として、第一に、嫌悪と恐怖の刺激を改善する必要がある。嫌悪は怒りや悲しみと混同されやすく、恐怖は驚きと混同されやすかった。これらの表情については、より典型的な顔面特徴を明示したプロンプトを用いて刺激を再生成したり、多数の刺激を生成した中から分類正答率の高い刺激を選別することが必要である。

第二に、悲しみ表情の強度を高める工夫が必要である。悲しみは分類成績では一定の水準を示したものの、強度評定では尺度中点を下回っており、より明確な悲しみ表情を生成する必要がある。具体的には、単に「悲しい表情」と指定するだけでなく、悲しみに特徴的な顔面動作をプロンプト上でより明示することが有効であると考えられる。たとえば、眉の内側が上がること、眉が中央に寄ることなど、悲しみ表情に特徴的な形態の手がかりを具体的に指定する必要がある。

第三に、中立表情や自然さに関する評価を追加する必要がある。本研究の強度評定課題では中立表情を対象に含めなかったため、中立表情がどの程度自然な無表情として知覚されるかは十分に検討できていない。また、生成 AI 画像は実写画像とは異なる人工的な特徴を含む可能性があるため、今後は表情カテゴリや強度だけでなく、自然さ、人工物感、魅力度、年齢印象などの基礎的情報も付与する必要がある。加えて、生成 AI によって作成された顔は類似した顔が多いため、類似した顔を除外することや、できるだけ多様な顔を作成するためのプロンプトの工夫なども、今後必要になるだろう。

第四に、既存の実写顔画像データベースとの直接比較が必要である。本研究では KRC データベースとの比較を行ったが、行った課題が完全に一致しているわけではないため、厳密な直接比較を行うことができなかった。特に、表情強度評定では、本研究で作成した生成 AI 画像の評定値が KRC データベースの値よりも全般的に低い傾向がみられた。しかし、この差が生成 AI 画像の表情強度の低さを反映しているのか、それとも評定課題の構造や評定文脈の違いによるものなのかは、本研究の比較だけでは判断できない。今後は、生成 AI 画像と実写顔画像を同一参加者、同一課題、同一手続きで評価し、生成 AI 顔画像が既存の実写顔画像と比較してどのような特徴や限界をもつのかを検討する必要がある。特に、生成 AI 画像において表情強度が全般的

に低く評価される傾向が再現されるのか、方法を揃えた直接比較によって確認することが重要である。

以上の課題を踏まえ、今後は、各刺激について分類正答率、表情強度評定値、誤分類の傾向などの妥当性情報を付与し、研究目的に応じて刺激を選択できる生成 AI 表情画像データベースとして整備していく必要がある。本研究は、生成 AI によって作成された顔表情画像が、実証的な妥当性評価を経ることで、心理学研究に利用可能な刺激セットとなりうることを示した。今後、刺激単位の評価情報を蓄積し、表情カテゴリや人物数を拡張していくことで、生成 AI を用いた顔画像データベースを、より汎用性と再利用性の高い研究資源として発展させることができると考えられる。

6 まとめ

本研究では、生成 AI を用いて日本人顔表情画像を作成し、表情分類課題と表情強度評定課題によって妥当性を検討した。その結果、生成 AI 表情画像は全体として一定の妥当性を有し、心理学実験刺激として利用できる可能性が示された。一方で、表情カテゴリや個別刺激によって妥当性には差があり、とくに嫌悪、恐怖、悲しみについては改善や選別が必要であることが明らかになった。今後、各刺激に分類成績、強度評定値、混同パターンなどの妥当性情報を付与することで、生成 AI を用いた日本人顔画像データベースを、より実用性の高い研究資源として整備し、将来的には幅広く利用可能な形で公開したいと考えている。

【参考文献】

- Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect “deepfake” images of human faces. *Journal of Cybersecurity*, 9(1), tyad011. <https://doi.org/10.1093/cybsec/tyad011>
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124–129. <https://doi.org/10.1037/h0030377>
- Fujimura, T., & Umemura, H. (2018). Development and validation of a facial expression database based on the dimensional and categorical model of emotions. *Cognition and Emotion*, 32(8), 1663–1670. <https://doi.org/10.1080/02699931.2017.1419936>
- Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion*, 22(6), 1094–1118. <https://doi.org/10.1080/02699930701626582>
- Han, S., Guo, Y., Zhou, X., Huang, J., Shen, L., & Luo, Y. (2023). A Chinese face dataset with dynamic expressions and diverse ages synthesized by deep learning. *Scientific Data*, 10, 878. <https://doi.org/10.1038/s41597-023-02701-2>
- 蒲池みゆき (2001). ATR 顔表情データベース(DB99)概要. ATR テクニカルレポート, TR-H-0305.
- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, 24(8), 1377–1388. <https://doi.org/10.1080/02699930903485076>
- Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science*, 34(12), 1390–1403. <https://doi.org/10.1177/09567976231207095>
- Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America*, 119(8), e2120481119. <https://doi.org/10.1073/pnas.2120481119>
- Roy-Charland, A., Perron, M., Beaudry, O., & Eady, K. (2014). Confusion of fear and surprise: A test of the perceptual-attentional limitation hypothesis with eye movement monitoring. *Cognition and Emotion*, 28(7), 1214–1222. <https://doi.org/10.1080/02699931.2013.878687>
- Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions:

- Judgments from untrained research participants. *Psychiatry Research*, 168(3), 242–249.
<https://doi.org/10.1016/j.psychres.2008.05.006>
- Ueda, Y., Nunoi, M., & Yoshikawa, S. (2019). Development and validation of the Kokoro Research Center (KRC) Facial Expression Database. *Psychologia*, 61(4), 221–240.
<https://doi.org/10.2117/psysoc.2019-A009>
- 渡邊伸行・鈴木竜太・吉田宏之・續木大介・番場あやの・Chandrasiri, N. P.・時田学・和田万紀・森島繁生・山田寛 (2007). 顔情報データベース FIND—日本人の顔画像データベース構築の試み— 感情心理学研究, 14(1), 39–53. <https://doi.org/10.4092/jsre.14.39>

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
画像生成 AI による心理学研究用顔画像データベースの作成——予備的検討——	日本感情心理学会第 33 回大会	2025. 10
画像生成 AI による心理学研究用顔画像データベースの作成——予備的検討(2)——	日本基礎心理学会第 44 回大会	2025. 12