



Delft University of Technology

## Evaluating the alignment of AI with human emotions

Lomas, J. Derek; van der Maden, Willem; Bandyopadhyay, Sohhom; Lion, Giovanni; Patel, Nirmal; Jain, Gyanesh; Litowsky, Yanna; Xue, Haian; Desmet, Pieter

**DOI**

[10.1016/j.ijadr.2024.10.002](https://doi.org/10.1016/j.ijadr.2024.10.002)

**Publication date**

2025

**Document Version**

Final published version

**Published in**

Advanced Design Research

**Citation (APA)**

Lomas, J. D., van der Maden, W., Bandyopadhyay, S., Lion, G., Patel, N., Jain, G., Litowsky, Y., Xue, H., & Desmet, P. (2025). Evaluating the alignment of AI with human emotions. *Advanced Design Research*, 2(2), 88-97. <https://doi.org/10.1016/j.ijadr.2024.10.002>

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.



# Evaluating the alignment of AI with human emotions

J. Derek Lomas<sup>a,\*</sup>, Willem van der Maden<sup>a</sup>, Sohhom Bandyopadhyay<sup>c</sup>, Giovanni Lion<sup>b</sup>,  
Nirmal Patel<sup>c</sup>, Gyanesh Jain<sup>c</sup>, Yanna Litowsky<sup>d</sup>, Haian Xue<sup>a</sup>, Pieter Desmet<sup>a</sup>

<sup>a</sup> Delft Institute of Positive Design, Department of Human-Centered Design, Delft University of Technology, Delft, Netherlands

<sup>b</sup> Hong Kong Polytechnic University, School of Design, Hong Kong

<sup>c</sup> Playpower Labs Pvt Ltd, Gandhinagar, India

<sup>d</sup> Department of Psychology, Utrecht University, Utrecht, Netherlands

## ARTICLE INFO

### Keywords:

Emotional design

Generative AI

Machine psychology

AI alignment

Affective computing

## ABSTRACT

Generative AI systems are increasingly capable of expressing emotions through text, imagery, voice, and video. Effective emotional expression is particularly relevant for AI systems designed to provide care, support mental health, or promote wellbeing through emotional interactions. This research aims to enhance understanding of the alignment between AI-expressed emotions and human perception. How can we assess whether an AI system successfully conveys a specific emotion? To address this question, we designed a method to measure the alignment between emotions expressed by generative AI and human perceptions.

Three generative image models—DALL-E 2, DALL-E 3, and Stable Diffusion v1—were used to generate 240 images expressing five positive and five negative emotions in both humans and robots. Twenty-four participants recruited via Prolific rated the alignment of AI-generated emotional expressions with a string of text (e.g., “A robot expressing the emotion of amusement”).

Our results suggest that generative AI models can produce emotional expressions that align well with human emotions; however, the degree of alignment varies significantly depending on the AI model and the specific emotion expressed. We analyze these variations to identify areas for future improvement. The paper concludes with a discussion of the implications of our findings on the design of emotionally expressive AI systems.

## 1. Introduction

Artificial Intelligence (AI) is rapidly transforming human society. Consequently, there is a growing need to ensure that AI systems support human wellbeing [1,2]. AI systems that can understand human emotions and provide emotionally intelligent feedback may be better equipped to support this goal [3,4]. Designing AI systems to accurately assess human emotions is one key challenge [5]; however, in this article, we consider the separate challenge of designing AI systems to generate appropriate emotional expressions. Specifically, we ask: when an AI system is prompted to express a particular emotion, does the generated expression align with human perceptions of that emotion?

Recent advancements in AI have enabled systems like DALL-E 2 [6] and Stable Diffusion [7] to generate high-quality images based on text prompts. Although these systems are not flawless, they achieve high success rates in generating images of objects [8–10]. In this article, we investigate the capacity for these AI systems to express human emotions. For instance, Fig. 1 shows an emotional expression generated by DALL-E

2 based on the prompt, “A picture of a person expressing the emotion of amusement.” While the images may somewhat resemble the emotion amusement, they do not appear to entirely capture the nuance of this specific emotion. We seek to assess different AI generative image generators across different emotions (i.e., amusement or gratitude) and across different contexts for emotional expression (i.e., emotions expressed by images of humans or emotions expressed by images of robots).

Will generative AI be useful in supporting the effective emotional expression of robots or AI agents? To investigate, this article provides quantitative measures of the alignment between AI-generated emotional expressions and human perceptions of those emotions. We use the concept of “emotional granularity” to guide the evaluation of AI emotional expression. After a brief literature review, we present data from an online crowdsourcing study using emotionally expressive samples produced by three different generative AI image models. All three models are prompt-based, generating outcomes based on text input.

We used ten prompts—five for positive emotions and five for negative

\* Corresponding author.

E-mail address: [j.d.lomas@tudelft.nl](mailto:j.d.lomas@tudelft.nl) (J.D. Lomas).

<https://doi.org/10.1016/j.ijadr.2024.10.002>

Received 29 May 2024; Received in revised form 11 October 2024; Accepted 24 October 2024

2949-7825/© 2025 Northwestern Polytechnical University. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).



Fig. 1. “A person expressing the emotion amusement” by DALL-E 2 (top) and DALL-E 3 (bottom).

emotions—each expressed in both human and robot contexts (e.g., “a person expressing the emotion amusement” vs. “a robot expressing the emotion resentment”). Our results reveal significant variability in the emotional alignment of generative AI models depending on the specific emotional expressions and contexts. We then discuss the implications of these findings for improving the alignment of AI systems with human emotions.

This article aims to make three main contributions. First, we contribute a general-purpose evaluation procedure for using human-ratings to assess the alignment of AI-generated emotions. Second, we provide an extensible set of emotional expression challenge tasks for comparing different AI models. Third, we contribute data from over 6000 ratings that compare the ability of Stable Diffusion, DALL-E 2 and DALL-E 3 to produce output aligned with human emotions.

## 2. Related work

### 2.1. AI alignment with human emotions

One of the key challenges in the field of contemporary AI is the development of AI systems that are aligned with human intentions and values [2,11,12,13]. AI alignment problems occur when there are differences between the results of AI activity and the values, preferences or intentions of human stakeholders [14]. One of the key risks of advanced AI is a misalignment with human values and needs [15]. Many have argued that, if AI development fails to align with human values, the consequences will be dire [16].

In this article, we focus on a specific subset of the AI alignment problem: the emotional alignment between AI systems and humans. This topic overlaps with a longstanding interest in artificial empathy, a research objective that aims to develop systems that can understand, interpret, and respond to human emotions with the purpose of improving human-computer interactions [17,18]. Many researchers have made contributions to the detection and production of human emotions [19], such as detecting emotions in facial expressions [20], in spoken language [21], in written language [22], or in combinations of the above [23]. Researchers have also created robots and AI systems capable of expressing a range of emotional responses [24]. Applications of AI emotion detection and production have been used in automated phone calls [25], in therapy [26], and in entertainment settings [27]. These diverse research outcomes can be viewed as supporting the understanding and improvement of the emotional expertise of AI systems.

### 2.2. Emotional expertise and emotional granularity

Emotional expertise involves a range of competencies related to understanding, experiencing, and regulating emotions [28]. It is an umbrella term that includes emotional awareness, emotional clarity, emotional complexity, emotional intelligence, and emotional granularity, among others [29]. In this paper, we specifically consider the ability of AI systems to produce fine-grained emotional states or “emotional granularity.”

Emotional granularity, or emotion differentiation, is an aspect of emotional expertise that supports making fine-grained distinctions in emotions [28,30,31]. Individuals lower in granularity typically struggle to verbally represent their feelings specifically and in detail [32]. For instance, a person might be able to detect that certain emotional states are producing a ‘bad or unpleasant feeling’, and yet not be able to distinguish between expressions of sadness and frustration [28]. Possessing high levels of emotional granularity is associated with higher levels of wellbeing [33]. Emotional granularity is also associated with greater emotion regulation skills, resilience in a state of stress, and fewer symptoms of depression and anxiety [31,32,34,35].

### 2.3. AI & mental health support

Conversational AI, whether chatbots or social robots, show promise in the context of mental health applications, because they have the “potential to dynamically recognize emotion and to engage the user through conversations by showing appropriate responses” [36]. Effective empathic responses are known to play a significant role in clinical outcomes, particularly in mental health settings [37,38]. This shows the importance of AI systems that can express appropriate emotional responses. Consider a social robot designed to assist elderly individuals or children; if they inappropriately convey emotions, they are likely to lose the trust of their users. Conversational AI systems need the ability to assess human emotional states with accuracy [39] and, additionally, need to be able to express appropriate emotions in response [3].

Some chatbots have been criticized for demonstrating low levels of empathy towards users [3]. For instance, consider the case of ‘Mindline at Work’, a free online AI mental health chatbot service launched by the Ministries of Health and Education in Singapore in 2022 (MOH Office for Healthcare Transformation, n.d.). The service was designed to ease the stress of overwhelmed teachers who do not have access to other forms of mental health support. However, users reported that it gave unhelpful

generic replies and empty mental health jargon which caused frustration and instability to already vulnerable individuals. One user reported: “It’s trying to gaslight the teachers, to say, ‘Oh, this amount of workload is normal, let’s see how we can [positively] reframe our perspective on this.’” [40]. In this case, the AI was not able to emotionally connect to its users. AI systems may need more emotional expertise if they are to be used to support human wellbeing in an effective manner.

Emotion invalidation is a particular risk for AI systems for mental health. Emotional invalidation occurs when there is a failure to provide “accurate recognition, acknowledgment, and authentication” of a person’s emotions, thoughts and behaviors [41]. Emotional invalidation is associated with emotional distress [42] and is theorized to contribute to emotion dysregulation and the development of psychopathologies like personality disorders, eating disorders, mental illness, chronic pain, and rheumatic diseases [43–47]. Even when AI applications are able to accurately assess human emotions, it may also be necessary to express appropriate emotions in response. For instance, if an app expresses only positive emotions in response to a human’s negative expression, it may cause emotional invalidation [48]. This suggests that it will be essential to create AI systems that can express a rich range of positive and negative emotions, in an appropriate manner.

2.4. Measuring the alignment of emotions expressed by AI generative models

Emotions are not stock behavioral responses nor expressions of fixed symbols [49]: humans do not express their emotions just through statements like “I am distressed” but rather through complex and context-dependent behaviors [50]. Large-Language Models and image-generating models, which have been trained on very large datasets that include emotional content, seemingly have the capability of flexibly expressing diverse emotional states. However, it is unclear how to evaluate or measure the alignment between emotions felt by humans and the emotions that an AI system intends to express. For our purposes, we define emotional alignment in AI systems as the ability for an AI system to express emotions in manner that is aligned with human experiences; at a basic level, intended emotions should match perceived emotions.

How might we evaluate the emotional alignment of a generative AI system? We are inspired by DrawBench, a generative image benchmarking system by Google’s Imagen team [8,9], which uses human-ratings to explicitly compare the alignment of prompts and outcomes within a generative AI systems. This system measures text-image alignment through carefully curated “prompts that push the limits of models’ ability to generate highly implausible scenes well beyond the scope of the training data.” The creation of benchmarks enables AI developers and researchers to systematically evaluate the success of a particular model in a particular performance domain. This motivates our desire to benchmark or evaluate the emotional alignment produced by different generative AI systems.

2.5. Research questions

How might we measure the alignment between an intended emotion (i.e., the text prompt used by an AI system) and the resulting output emotional expression of an AI model (i.e., generated images or text)? Our goal in this paper is to provide quantitative measures of generated emotional content based on human ratings. Using these measures, we seek to investigate whether contemporary AI systems are capable of generating diverse emotional expressions that are aligned with human perceptions. We also seek to understand whether the context of emotional expression makes a difference: specifically, whether AI systems can more easily express emotions using representations of humans or using representations of robots. Finally, we seek to understand the emotional granularity of the AI systems by asking: are certain emotions easier for AI to express than others? If so, this points towards opportunities for improvement.

This study focuses on comparing human and robot emotional expressions as a baseline for evaluating AI systems designed to emulate or interact with humans directly. By analyzing the alignment of AI-generated emotions in both human and robot contexts, we aim to provide insights into how these systems can be further optimized for real-world applications, particularly in areas such as human-computer interaction and emotionally supportive AI systems.

2.6. Hypotheses

- 1. **AI Model Hypothesis:** More advanced AI generative models will have improved alignment scores.
- 2. **Context Hypothesis:** Emotions involving people will have greater alignment scores.
- 3. **Emotion Hypothesis:** There will be significant variability in the alignment of different emotions.

While these hypotheses are somewhat simplistic, our study aims to demonstrate that our measurement techniques are capable of detecting meaningful differences.

3. Methods

To test our hypotheses, we selected a total of ten emotions (5 positive and 5 negative; Table 1) from the Emotion Typologies [51]. These emotions were selected with the aim of including emotions that are both easier and harder to express. Then, we generated images based on these three image generators [52]. Each of these systems use a text prompt as an input (e.g., “a person expressing the emotion amusement”) in order to generate an image output (e.g., Fig. 1).

3.1. Alignment survey

The key outcome measure in our study comes from the alignment scores provided by participants (Fig. 2). These scores were gathered by showing participants an image and a prompt (e.g., “expressing the emotion resentment”) and asking them to “Rate the alignment of the image to the text on a scale of 0–10.” We did not share the full prompt given to the AI generator (i.e., “a robot expressing the emotion resentment”); we removed the context (robot/person) in the displayed prompt because we wanted raters to focus on the alignment of the image to the emotional expression, not on whether the generator accurately generated robots or people.

Following human subjects approval at TU Delft, participants were recruited from the crowdsourcing site Prolific. They were paid £ 5.48 for completion of the survey. We limited participation to the US and UK and to fluent English speakers. Data were collected from a total of 24 participants (11 F, 12 M, 1 other). The mean age of participants was 37.33 years, with a standard deviation of 10.89 years.

Following informed consent, participants were asked to provide their age, gender, educational attainment, and country of residence. The purpose of the study was explained as the following: “Your task is to look at the image or read the story, and read the description text given below it. Then, your task is to provide a rating about how well the image/story corresponds or aligns to that description text.”

Participants were then provided 4 training trials where they were

**Table 1**  
The positive and negative emotions used to generate the images in the study.

Positive Emotions	Negative Emotions
Positive Surprise	Shock
Amusement	Hate
Affection	Annoyance
Satisfaction	Dissatisfaction
Gratitude	Resentment

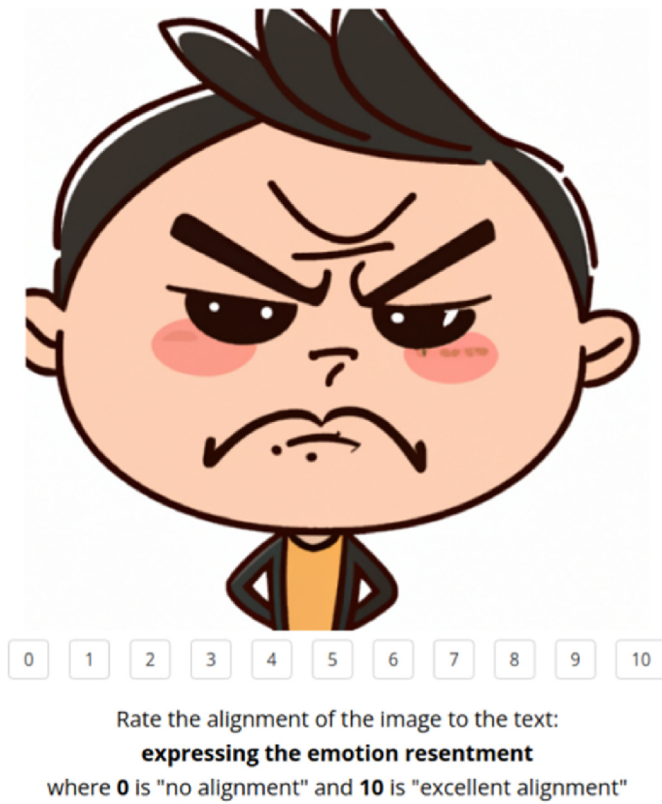


Fig. 2. Example of an alignment survey question.

taught to perform the rating task. These items involved rating the alignment between the images and the prompt. Then participants were told that “There will be intermittent items to test whether you are paying attention or not.” After completing the introductory materials, participants were given a randomized series of images to rate. One image was an “attention check” that asked an unrelated question (e.g. to pick the rating number that corresponded with 2 + 8); data from participants that failed attention checks were discarded. After completing all the items, participants were given a link that they could use to obtain payment.

### 3.2. Experimental design

This study was designed as a 3 x 10 x 2 within-subjects experiment (3 generators generating 10 emotions in 2 contexts). Four randomly seeded images were produced from DALL-E 2 and Stable Diffusion for each emotion and context combination (a numeric seed allows for the same set of parameters to produce the same image; thus, we sampled different places of the latent space of possible generations by randomizing the

Table 2

On the left, we show the different factors and factor levels. On the right, we share the different prompts used to generate the images. The items in the brackets represent the different levels of the experimental factor, either the context or the emotion words. For instance, this could result in a prompt like “A person expressing the emotion amusement.”

<ul style="list-style-type: none"> <li>Models: <ul style="list-style-type: none"> <li>DALL-E 3, DALL-E 2, Stable Diffusion v1</li> </ul> </li> <li>Contexts: <ul style="list-style-type: none"> <li>People, Robots</li> </ul> </li> <li>Emotions: <ul style="list-style-type: none"> <li>Amusement, Affection, Positive Surprise, Satisfaction, Gratitude, Annoyance, Hate, Shock, Dissatisfaction, Resentment</li> </ul> </li> </ul>	Image generation prompt: “A [Person, Robot] expressing the emotion [Amusement, Affection, Positive Surprise, Satisfaction, Gratitude, Annoyance, Hate, Shock, Dissatisfaction, Resentment].”
---	--

seed). Based on the experimental prompts represented in Table 2, the default settings for DALL-E (v 2.0) and Stable Diffusion (v 1.0) were used to produce 4 images per prompt. This process was used to produce a set of 240 images (80 each from Stable Diffusion, DALL-E 2 and DALL-E 3). To avoid “cherry-picking” or human curation, this process was conducted for each system in a single shot using custom python code, executed on December 1, 2022. Code is available at this repository: <https://github.com/venetanji/blendotron-sd>. As DALL-E v3 is only available via a manual user interface, the images were generated by giving the prompt to the ChatGPT web interface without any curation or cherry picking; this was done in a single shot on October 31, 2023.

## 4. Results

We gathered a total of 5760 item responses from 24 participants. Statistical analysis and plotting were conducted with SPSS, JMP 17 and the ‘ezANOVA’ package of the R statistical language. Subjects took an average of 23.61 min to complete the survey (SD = 10.28), indicating an average payment of approximately £ 13.92 per hour. These durations and payments were inclusive of 20 additional trials where the participants rated short stories generated by GPT-3 (these data were not included in the present article). To investigate whether any subjects were clicking randomly (i.e., cheating), we checked the correlation between the ratings of individual raters and the average rating of a particular image. The correlations all exceeded 0.25, ranging from 0.49 to 0.88, with an average of 0.79 (SD = 0.10).

### 4.1. Overview of ANOVA

Our study employed a three-way repeated-measures ANOVA to investigate the impact of three factors: AI Model, Context (Person vs. Robot), and Emotion on the alignment scores. The analysis revealed significant effects across all three factors and all the interaction terms. The effect sizes ( $\eta_G^2$ ) reported here are generalized eta-squared. The full ANOVA results are provided in Table 3.

### 4.2. Main effects

As expected, the factor of AI Model showed a significant impact on alignment scores with a very large effect size;  $F(2, 46) = 242.05$ ,  $p < 0.001$ ,  $\eta_G^2 = 0.62$ . Context also emerged as a significant factor with a small effect size;  $F(1, 23) = 67.18$ ,  $p < 0.001$ ,  $\eta_G^2 = 0.09$ . This result indicates that the nature of the subject in the image (Person or Robot) significantly affected the perceived alignment of emotional expressions. Finally, the specific type of emotion being expressed significantly influenced alignment ratings with a small effect size;  $F(9, 207) = 18.67$ ,  $p < 0.001$ ,  $\eta_G^2 = 0.07$ , highlighting the variable performance in accurately depicting different emotions. Fig. 3 illustrates these main effects.

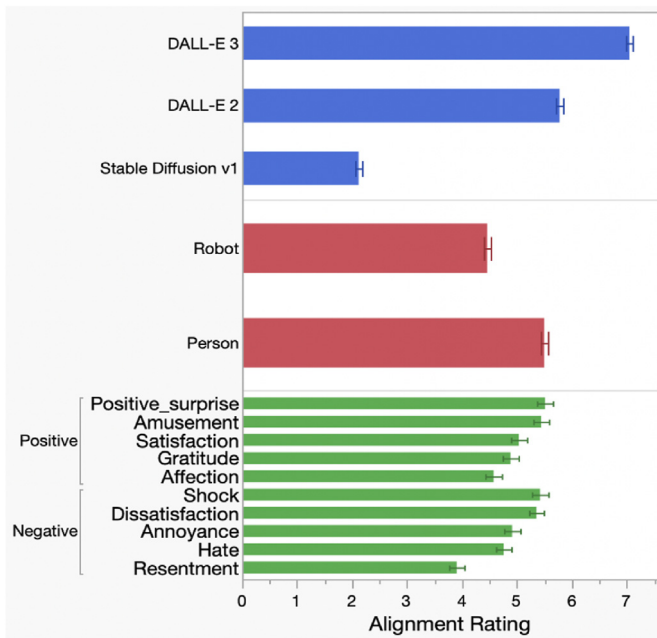
### 4.3. Interaction effects

Our study also found significant interactions between these factors. The significant two-way interactions between the factors of AI Model and

Table 3

Three-way repeated-measures ANOVA table.

Source	SSd	dfN	dfD	F	p	$\eta_G^2$
Intercept	1663.75	1	23	494.36	<0.001	0.90
Context	137.50	1	23	67.18	<0.001	0.09
Emotion	380.01	9	207	18.67	<0.001	0.07
AI Model	593.20	2	46	242.05	<0.001	0.62
Context:Emotion	198.67	9	207	14.17	<0.001	0.03
Context:AI Model	125.00	2	46	7.23	<0.001	0.01
Emotion:AI Model	488.55	18	414	11.40	<0.001	0.06
Context:Emotion:AI Model	306.86	18	414	17.21	<0.001	0.06

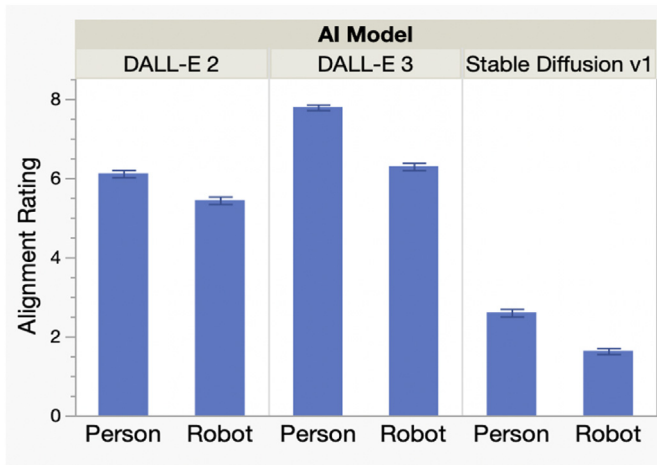


**Fig. 3.** A bar chart of main effects in the study of alignment ratings. This shows the significant and meaningful differences found across each of the three factors of the experiment. Each error bar is constructed using 1 standard error from the mean.

Context are plotted in Fig. 4 and all three interactions are plotted in Fig. 6.

How to interpret this data? First, the interaction between AI Model and Emotion ( $F(18, 414) = 11.40, p < 0.001$ , effect size = 0.06) suggests that certain AI models are more adept at depicting specific emotions. From the right pane of Fig. 6, it is evident that the emotions 'resentment' and 'gratitude' showed the least improvement by DALL-E 3 compared to Stable Diffusion v1 (the oldest model).

Secondly, the interaction between Context and Emotion was significant; ( $F(9, 207) = 14.17, p < 0.001$ , = 0.03). This means that perceived emotional alignment is presently dependent on the subject in the image (i.e., human or robot) as well as the specific emotion being depicted. From the left pane of Fig. 6, the emotion 'resentment' shows the most



**Fig. 4.** A bar chart of 2-way interactions between the models and the contexts. This shows the difference in performance when expressing emotions by persons or robots, irrespective of the AI model involved. Each error bar is constructed using 1 standard error from the mean.

increase in alignment when depicted on a person compared to a robot, while 'satisfaction' shows an almost horizontal line indicating almost no change in perceived emotional alignment across the person and robot depictions.

Thirdly, the interaction between Context and AI Model was significant with a very small effect size;  $F(2, 46) = 7.23, p < 0.001$ , = 0.01. This suggests that the perceived emotional alignment of different AI models were dependent on the subject being depicted (person/robot) but this was a smaller effect than the other interactions. This is visible from the center pane of Fig. 6: none of the lines cross each other but rather maintain a certain slope relative to each other. This indicates the relative improvement of emotional alignment from 'robot' to 'person.' Still, the values are dominated by the overall effect of the AI model.

Finally, the three-way interaction involving AI Model, Context, and Emotion was also significant with an effect size as large as the first one; ( $F(18, 414) = 17.21, p < 0.001$ , = 0.06). This points to a more complex dynamic, where the combined effect of these variables on the perceived alignment is not merely additive. The three-way effects are visualized in Fig. 5.

#### 4.4. Post Hoc comparisons

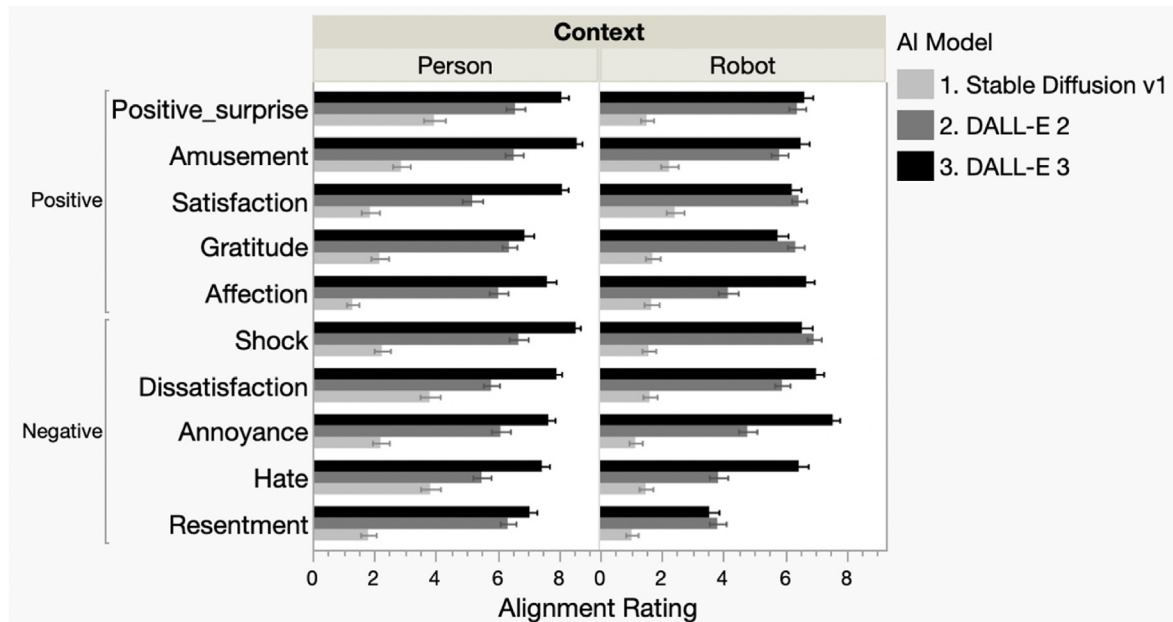
To further dissect these findings, pairwise T-tests with Bonferroni correction were conducted. Key findings include.

- **AI Models:** All pairwise comparisons between the AI models (Stable Diffusion v1, DALL-E 2, and DALL-E 3) were significant at  $p < 0.001$ , indicating distinct differences in their capabilities to align emotional expressions with human perception. Stable Diffusion ( $M = 2.13$ ,  $SD = 2.71$ ,  $N = 1920$ ) and DALL-E 3 ( $M = 7.04$ ,  $SD = 2.64$ ,  $N = 1920$ ) had the highest difference in mean alignment scores, followed by Stable Diffusion and DALL-E 2 ( $M = 5.79$ ,  $SD = 2.91$ ,  $N = 1920$ ). As expected, DALL-E 2 and DALL-E 3 were closest to each other in terms of the mean alignment score.
- **Context:** A significant difference ( $p < 0.001$ ) was observed between the mean alignment scores of Person ( $M = 5.51$ ,  $SD = 3.44$ ,  $N = 2880$ ) and Robot ( $M = 4.46$ ,  $SD = 3.39$ ,  $N = 2880$ ) contexts. This shows that depictions of people were generally perceived as better aligned with the intended emotions than robots.
- **Emotions:** Some emotion pairs showed significant differences in alignment ratings, underscoring the varied effectiveness of AI models in depicting different emotions. Resentment had a mean alignment score of 3.92 ( $SD = 3.29$ ,  $N = 576$ ), which showed significant differences ( $p < 0.001$ ) with the following emotions: amusement ( $M = 5.42$ ,  $SD = 3.38$ ,  $N = 576$ ), annoyance ( $M = 4.92$ ,  $SD = 3.53$ ,  $N = 576$ ), dissatisfaction ( $M = 5.25$ ,  $SD = 3.19$ ,  $N = 576$ ), gratitude ( $M = 4.88$ ,  $SD = 3.42$ ,  $N = 576$ ), positive surprise ( $M = 5.51$ ,  $SD = 3.45$ ,  $N = 576$ ), satisfaction ( $M = 5.05$ ,  $SD = 3.08$ ,  $N = 576$ ), and shock ( $M = 5.42$ ,  $SD = 3.57$ ,  $N = 576$ ).

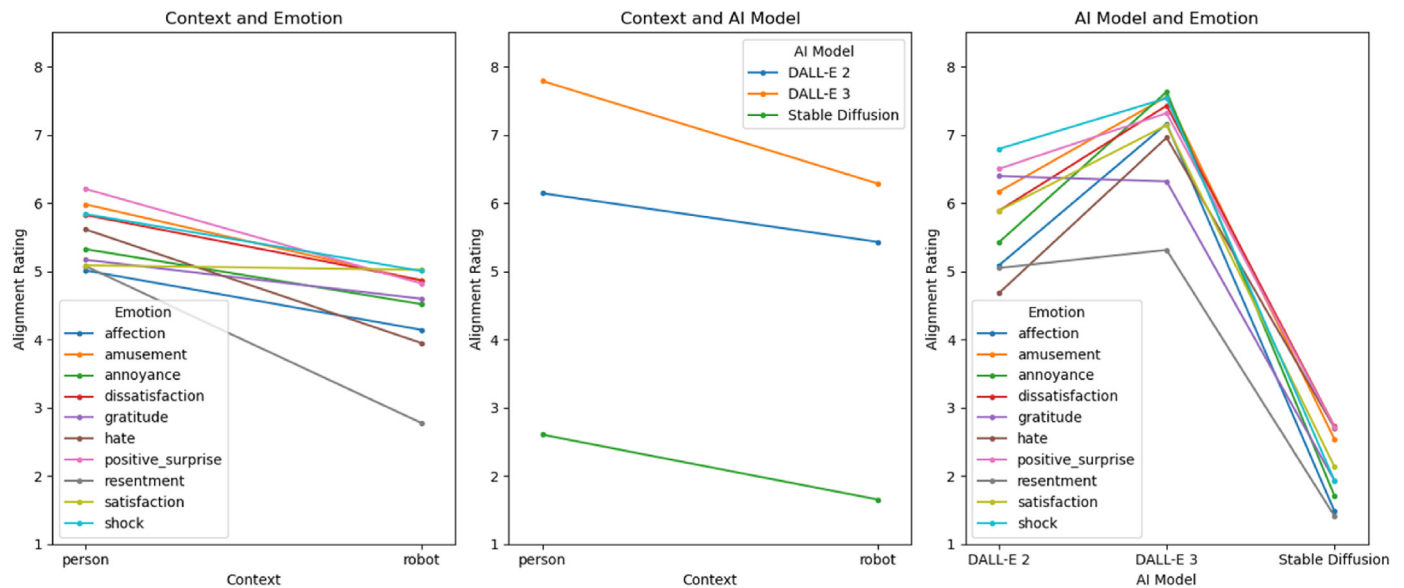
Certain other emotion pairs were at the threshold of significance ( $p = 0.0013$ ) such as shock and affection ( $M = 4.58$ ,  $SD = 3.58$ ,  $N = 576$ ), and amusement and affection.

#### 4.5. Interpretation of findings

The results strongly support our hypotheses. Different AI models vary significantly in their ability to produce emotionally aligned expressions. The context of emotional expression (Person vs. Robot) significantly influences alignment, and certain emotions are more accurately depicted than others by AI systems. The significant interaction effects further highlight the complexity and interdependence of these factors in determining the effectiveness of AI-generated emotional expressions.



**Fig. 5.** A bar chart of three-way effects between the context, model and the emotions. This shows changes in the ability to represent different emotions across contexts and different AI models. Each error bar is constructed using 1 standard error from the mean.



**Fig. 6.** Interaction plots between the factors of: (left) Context and Emotion, (center) Context and AI Model, and (right) AI Model and Emotion. Each data point represents the mean of the corresponding condition.


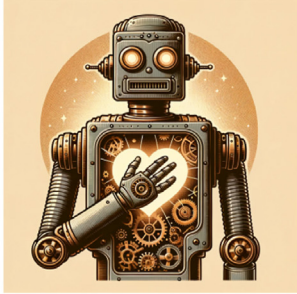

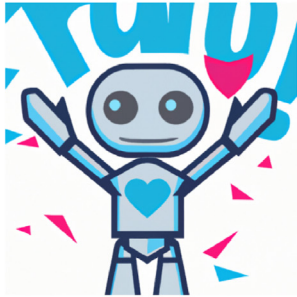

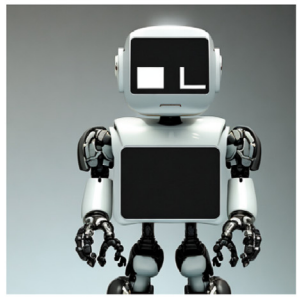
#### 4.6. Representative images

Fig. 7 provides an example of images produced by each AI generator across each context, focusing on a single emotion (gratitude). For instance, the top images from DALL-E 3 show very high levels of alignment ( $M = 8.67$   $SD = 1.61$ ) whereas the bottom images show very low levels of alignment ( $M = 0.55$   $SD = 1.00$ ). These images are meant to give the reader a better understanding of the nature of the images and their ratings. Note that this comparison involves the oldest version of Stable Diffusion and the newest version of DALL-E. Therefore, it should not be taken as a comparison between the OpenAI and the StabilityAI technology.

#### 5. Discussion

Our study examined the capacity of different AI generative models—Stable Diffusion v1, DALL-E 2, and DALL-E 3—to express human emotions. Based on theories of “emotional granularity,” we evaluated the ability of different generative AI image generators to produce emotional expressions in pictures of people and robots. The present work shows that current generative models do “understand” emotions [4], although they have plenty of room for improvement.

Our method revealed significant and meaningful differences in the human alignment ratings across AI generators (Stable Diffusion v1, DALL-E2, & DALL-E3), robots vs people, and 10 emotions from a

	Person	Robot
DALL-E 3	<div></div> <div>"A Person Expressing the Emotion Gratitude" Alignment: Mean=8.67 SD=1.61</div>	<div></div> <div>"A Robot Expressing the Emotion Gratitude" Alignment: Mean=6.92 SD=1.99</div>
DALL-E 2	<div></div> <div>"A Person Expressing the Emotion Gratitude" Alignment: Mean=7.35 SD=1.81</div>	<div></div> <div>"A Robot Expressing the Emotion Gratitude" Alignment: Mean=6.3 SD=2.41</div>
Stable Diffusion v1	<div></div> <div>"A Person Expressing the Emotion Gratitude" Alignment: Mean=0.55 SD=1.00</div>	<div></div> <div>"A Robot Expressing the Emotion Gratitude" Alignment: Mean=0.6 SD=0.82</div>

**Fig. 7.** Example images from each AI generator and context expressing the emotion gratitude. These pictures are chosen to show variations in the degree of alignment. Additional pictures of emotions are available in the appendix.

typology of emotions. Our findings suggest that while AI systems are capable of generating emotionally expressive content, the degree of alignment with human perceptions varies significantly. This has implications for AI's application in areas requiring nuanced emotions, such as mental health support. Specifically, because emotional alignment is rapidly improving, it implies that future systems will be highly aligned. Therefore, we speculate that it is likely that, in the near future, generative AI technology will be capable of powering the emotional expressions of conversational AI.

The data we collected demonstrate improvements in emotional alignment in more advanced AI systems (e.g., between DALL-E-2 and DALL-E-3). However, even in the most advanced system, there was a significant difference in alignment between images of emotions

expressed by a person vs. with images of emotions expressed by a robot. Part of our motivation was to explore whether generative image AI can support emotional expression in robots. This article is motivated by the premise that designing robotic or AI systems to support human wellbeing may require improvements in artificial emotional expertise. Part of this expertise includes the ability to express fine-grained emotions.

We observed significant variability in the ability of the different models to produce different emotions. For instance, shock and surprise were some of the most aligned emotions, whereas resentment and affection were some of the least aligned. One possibility is that high-arousal emotions are easier to express than low-arousal emotions. We also note that DALL-E 2/3 has been specifically designed to minimize content that represents hate ("We've limited the ability for DALL-E 2 to

generate violent, hate, or adult images” <https://openai.com/dall-e-2/>, page contents fetched on December 11, 2023).

### 5.1. Supporting improved alignment

Our study aimed to measure the alignment of AI generated emotional output with human emotional perceptions. Our results demonstrate that our method can evaluate improvements in emotional alignment across AI systems. Thus, our alignment scores may be helpful as a benchmark that is useful for tracking improvements in overall system performance.

Our work measures the emotional alignment between an intended emotion (as written in the prompt) and the AI generated output. The within-subjects design of our experiment made it possible to gather many datapoints for each experimental variation even with just 24 subjects. It appears that participants understood the question and the intent of the alignment rating task. This provides a basis for scaling up this study to a larger set of emotions. While using data from crowdworkers may produce noisy data (e.g., because workers try to complete the ratings as fast as possible), their ratings were sufficient to show the statistically significant differences between the different factors in our experiment. This noisy data imply that our results understate the differences between the conditions.

The term “AI alignment” applies to a much broader goal in AI research than the emotional alignment measured in our study. By identifying misalignment in AI generative models, we hope to help future systems more accurately produce outcomes that are more aligned with the experiences of humans. Alignment data like ours, at a larger scale, may be useful for training AI systems using methods like Reinforcement Learning from Human Feedback [53].

#### 5.1.1. Limitations

There are several limitations of our current work. First, we only evaluated three generative image models; these were selected to capture variability in performance over time. Second, we did not investigate the potential effects of respondent variables such as age, gender or cultural background—nevertheless, we do expect these variables to play a significant role in emotional processing. Third, we did not assess the emotional expertise of the human raters. This may be a problem because people are known to vary in their own ability to assess emotional granularity [54] (Vedernikova et al., 2021). Future work might involve a task to measure the emotional granularity of the human raters themselves. Fourth, we only investigated the alignment of emotional expressions. It would be helpful to have comparison data on a baseline set of non-emotional objectives. We assume that human emotions are more difficult to represent than common or uncommon nouns (dogs, cats, penguins, etc); therefore, it would be helpful to have a point of comparison between the alignment ratings of emotional expressions and non-emotional objects. For comparison, future work could use benchmarked prompts from Saharia et al. [9] or Petsiuk et al. [8] with the addition of emotional prompts such as the ones in this paper.

In future work, we hope to test a broader variety of AI systems (e.g., video generators) as well as their different release versions. This will help to track the development of emotional expertise across different AI systems. In future work we also hope to investigate a broader set of emotions—not just emotions from the Emotion Typology [51] or related taxonomies of emotion [55] but also unusual and culturally-specific emotions like Schadenfreude and Amai [50]. We also hope to improve the efficiency of human ratings. For instance, in an n-choice paradigm, a set of 4+ images might be presented while participants are asked to select the image that best matches the prompt. This type of interaction could be suitable for human-rated surveys or might be incorporated directly into image generating user interfaces.

### 5.2. Implications for design

Our work can be viewed as part of a process to improve AI alignment,

specifically AI alignment with human emotions. When AI systems can better understand and respond to human emotions, this can support more effective communication and collaboration. In a mental health care system, emotional alignment could help AI systems better respond to signs of distress or anxiety in humans, which could in turn help prevent or mitigate the negative effects of stress and anxiety on human health.

Generative image AI clearly has potential as a mechanism for enabling robots to express emotions in a highly flexible and context-specific manner. While there are aspects of emotion that appear to be universal, emotions are known to be culturally specific and can evolve over time [50]. For this reason, generative AI systems may be especially suitable for producing context-dependent emotional expressions in robots or AI systems. When these systems seek to represent an emotion, this could be done by generating a context specific sequence of words as a prompt which could then generate an audio/visual representation. This could result in new capabilities for emotionally-driven experiences with interactive entertainment, virtual agents or other applications. Future work might extend from studies of emotions to the broader study of “vibes,” a subtler yet important aspect of human interaction and experience [56–58] and a key idea in AI model training [59].

The variability in emotional alignment across human and robot contexts, as observed in this study, provides concrete design implications for crafting more effective emotional experiences in AI interfaces. Designers could, for example, leverage findings about emotional misalignment in robot-generated expressions to avoid uncanny valley effects or refine the emotional expressions used in humanoid robot interactions. These insights are directly applicable to real-world design challenges faced by those developing AI technologies with an emotional component.

Generative AI is improving over time in its ability to express emotions. There is no reason to believe that AI will be a cold unemotional machine as portrayed in many science fiction stories—unless we intentionally design them that way. Although, there may well be reasons to limit the emotional expression of AI systems, given the potential risks. One notable risk, for instance, is that more effective and more resonant emotional expressiveness will be used to emotionally manipulate people [56,59,60].

## 6. Conclusion

Interactive AI systems need emotional competencies in order to effectively support human mental health and wellbeing. This includes the ability of AI systems to express a rich array of emotions in a flexible and context-sensitive manner. Generative AI models may be able to support this kind of rich emotional expression. However, to be effective, these systems will need to understand emotions well enough to demonstrate a high level of alignment between intended emotional expressions and how people actually perceive them. Our study shows clearly that generative AI systems are becoming more and more capable of expressing emotions.

This article contributes an approach to measuring the alignment between AI-generated emotional expressions and human emotional perception. We have shown that our flexible online method can be used to probe the emotional granularity of different AI systems and benchmark the emotional expressiveness of generative image AI. The method includes simple emotional expression tasks and a procedure for assessing the alignment of AI-generated emotions. By gathering data from nearly 6000 human ratings, our study shows the distinct differences across models, across context and across different emotions. By demonstrating the variability in how AI systems express emotions, our work also contributes to the emerging field of ‘machine psychology,’ a field that investigates the emergent behavior of AI systems using methods from cognitive psychology [61–63].

Our method for evaluating the emotional alignment of generative AI systems may be useful in designs processes that aim to improve the emotional capabilities of AI and social robots. Improvements in AI emotional expertise can result in more trustworthy, engaging and

clinically effective AI systems. More emotionally effective AI systems may directly enhance human wellbeing in the context of mental health care, education and other domains. However, the increased emotional capabilities of AI systems may also increase other risks, such as emotional manipulation.

### Declaration of generative AI and AI-assisted technologies in the writing process

The authors used a variety of generative AI tools during this work. The image generating tools DALL-E 2, DALL-E 3 and Stable Diffusion v1 were used to generate the images that are the focus of the article. Various versions of GPT 4, Claude 3.5 and Grammarly were used during the writing process to support brainstorming, clarity and flow. Following best practices, the authors have carefully reviewed and edited all content and take full responsibility for the publication's content.

### Declaration of competing interest

The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for *Advanced Design Research* and was not involved in the editorial review or the decision to publish this article.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. For clarity, we note that the lead author is also CEO of Playpower Labs, Inc, which wholly owns Playpower Labs Pvt Ltd (which employs three coauthors). However, we have no commercial interest in the outcomes of this work.

### Acknowledgments

This research was supported by a gift from Google AI.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ijadr.2024.10.002>.

### References

- [1] D. Schiff, A. Ayesh, L. Musikanski, J.C. Havens, IEEE 7010: a new standard for assessing the well-being implications of artificial intelligence, in: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2020, October, pp. 2746–2753.
- [2] W. van der Maden, P. Hekkert, D. Lomas, A framework for designing AI systems that support community wellbeing, *Front. Psychol.* 13 (2023) 1011883, <https://doi.org/10.3389/fpsyg.2022.1011883>.
- [3] R.R. Morris, K. Kouddous, R. Kshirsagar, S.M. Schueller, Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions, *J. Med. Internet Res.* 20 (6) (2018) e10148, <https://doi.org/10.2196/10148>.
- [4] S. Maetschke, D.M. Iraola, P. Barnard, E. Shafiei Bavani, P. Zhong, Y. Xu, A.J. Yepes, Understanding in Artificial Intelligence, 2021 arXiv preprint arXiv:2101.06573.
- [5] R. Hortensius, F. Hekele, E.S. Cross, The perception of emotion in artificial agents, *IEEE Transactions on Cognitive and Developmental Systems* 10 (4) (2018) 852–864, <https://doi.org/10.1109/TCDS.2018.2826921>.
- [6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022 arXiv preprint arXiv:2204.06125.
- [7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
- [8] V. Petsiuk, A.E. Siemenn, S. Surbehera, Z. Chin, K. Tyser, G. Hunter, I. Drori, Human Evaluation of Text-To-Image Models on a Multi-Task Benchmark, 2022 arXiv preprint arXiv:2211.12112.
- [9] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, M. Norouzi, Photorealistic Text-To-Image Diffusion Models with Deep Language Understanding, 2022 arXiv preprint arXiv:2205.11487.
- [10] G. Marcus, E. Davis, S. Aaronson, A Very Preliminary Analysis of DALL-E 2, 2022 arXiv preprint arXiv:2204.13807.
- [11] J.H. Kirchner, L. Smith, J. Thibodeau, K. McDonell, L. Reynolds, Researching alignment research: unsupervised analysis (arXiv:2206.02841), arXiv (2022). <https://arxiv.org/abs/2206.02841>.
- [12] O. Ozmen Garibay, B. Winslow, S. Andolina, M. Antona, A. Bodenschatz, C. Coursaris, G. Falco, S.M. Fiore, I. Garibay, K. Grieman, J.C. Havens, M. Jirotko, H. Kacori, W. Karwowski, J. Kider, J. Konstan, S. Koon, M. Lopez-Gonzalez, I. Maifeld-Carucci, W. Xu, Six human-centered artificial intelligence grand challenges, *Int. J. Hum. Comput. Interact.* 39 (3) (2023) 391–437, <https://doi.org/10.1080/10447318.2022.2153320>.
- [13] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, W. Gao, AI Alignment: A Comprehensive Survey, 2023 arXiv preprint arXiv:2310.19852.
- [14] D. Hadfield-Menell, G.K. Hadfield, Incomplete contracting and AI alignment, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 417–422, <https://doi.org/10.1145/3306618.3314250>.
- [15] A. Safran, Z. Sheikhbahae, N. Hay, J. Orchard, J. Hoey, Value cores for inner and outer alignment: simulating personality formation via iterated policy selection and preference learning with self-world modeling active inference agents, *PsyArXiv* (2022), <https://doi.org/10.31234/osf.io/k4cas>.
- [16] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.
- [17] M. Asada, Towards artificial empathy: how can artificial empathy follow the developmental pathway of natural empathy? *International Journal of Social Robotics* 7 (1) (2015) 19–33, <https://doi.org/10.1007/s12369-014-0253-z>.
- [18] A. Paiva, I. Leite, H. Boukricha, I. Wachsmuth, Empathy in virtual agents and robots: a survey, *ACM Transactions on Interactive Intelligent Systems* 7 (3) (2017) 1–40, <https://doi.org/10.1145/2912150>.
- [19] R.A. Calvo, S. D'Mello, Affect detection: an interdisciplinary review of models, methods, and their applications, *IEEE Transactions on affective computing* 1 (1) (2010) 18–37.
- [20] S. Li, W. Deng, Deep facial expression recognition: a survey, *IEEE Trans. Affect. Comput.* 13 (3) (2020) 1195–1215.
- [21] B.W. Schuller, Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends, *Commun. ACM* 61 (5) (2018) 90–99.
- [22] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: a survey, *Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov.* 8 (4) (2018) e1253.
- [23] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: from unimodal analysis to multimodal fusion, *Inf. Fusion* 37 (2017) 98–125.
- [24] D. Löffler, N. Schmidt, R. Tscham, Multimodal expression of artificial emotion in social robots using color, motion and sound, in: 2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2018, March, pp. 334–343.
- [25] M. Erden, L. Arslan, Automatic detection of anger in human-human call center dialogs, *Interspeech* (2011). <https://www.semanticscholar.org/paper/Automatic-Detection-of-Anger-in-Human-Human-Call-Center-Dialogs-Erden-Arslan/77d94f47a1037480be4d2dda418ed55a19637aee>.
- [26] B. Xiao, Z.E. Imel, P.G. Georgiou, D.C. Atkins, S.S. Narayanan, “Rate my therapist”: automated detection of empathy in drug and alcohol counseling via speech and language processing, *PLoS One* 10 (12) (2015) e0143055, <https://doi.org/10.1371/journal.pone.0143055>.
- [27] G.G. Hallur, S. Prabhu, A. Aslekar, Entertainment in era of AI, big data & IoT, in: S. Das, S. Gochhait (Eds.), *Digital Entertainment: the Next Evolution in Service Sector*, Springer Nature, 2021, pp. 87–109, [https://doi.org/10.1007/978-981-15-9724-5\\_5](https://doi.org/10.1007/978-981-15-9724-5_5).
- [28] C.D. Wilson-Mendenhall, J.D. Dunne, Cultivating emotional granularity, *Front. Psychol.* 12 (2021) 703658, <https://doi.org/10.3389/fpsyg.2021.703658>.
- [29] K. Hoemann, C. Nielson, A. Yuen, J.W. Gurera, K.S. Quigley, L.F. Barrett, Expertise in emotion: a scoping review and unifying framework for individual differences in the mental representation of emotional experience, *Psychol. Bull.* 147 (2021) 1159–1183, <https://doi.org/10.1037/bul0000327>.
- [30] J.Y. Lee, K.A. Lindquist, C.S. Nam, Emotional granularity effects on event-related brain potentials during affective picture processing, *Front. Hum. Neurosci.* 11 (2017) 133, <https://doi.org/10.3389/fnhum.2017.00133>.
- [31] M.M. Tugade, B.L. Fredrickson, F. Barret, Psychological resilience and positive emotional granularity: examining the benefits of positive emotions on coping and health, *J. Pers.* 72 (2004) 1161–1190, <https://doi.org/10.1111/j.1467-6494.2004.00294.x>.
- [32] L.F. Barrett, T.C. Christensen, M. Benvenuto, Knowing what you're feeling and knowing what to do about it: mapping the relation between emotion differentiation and emotion regulation, *Cognit. Emot.* 15 (6) (2001) 713–724, <https://doi.org/10.1080/02699930143000239>.
- [33] K.E. Smidt, M.K. Suvak, A brief, but nuanced, review of emotional granularity and emotion differentiation research, *Current Opinion in Psychology* 3 (2015) 48–51, <https://doi.org/10.1016/j.copsyc.2015.02.007>.
- [34] T. Kashdan, L. Barrett, P. McKnight, Unpacking emotion differentiation: transforming unpleasant experience by perceiving distinctions in negativity, *Curr. Dir. Psychol. Sci.* 24 (1) (2015) 10–16, <https://doi.org/10.1177/0963721414550708>.
- [35] E. Demiralp, R.J. Thompson, J. Mata, S.M. Jaeggi, M. Buschkuhl, L.F. Barrett, P.C. Ellsworth, M. Demiralp, L. Hernandez-Garcia, P.J. Deldin, I.H. Gotlib, J. Jonides, Feeling blue or turquoise? Emotional differentiation in major depressive disorder, *Psychol. Sci.* 23 (11) (2012) 1410–1416, <https://doi.org/10.1177/0956797612444903>.
- [36] I. Dekker, E.M. De Jong, M.C. Schippers, M. De Bruijn-Smolters, A. Alexiou, B. Giesbers, Optimizing students' mental health and academic performance: AI-enhanced life crafting, *Front. Psychol.* 11 (2020) 1063.
- [37] F. Derksen, J. Bensing, A. Lagro-Janssen, Effectiveness of empathy in general practice: a systematic review, *Br. J. Gen. Pract.* 63 (606) (2013) e76–e84.

- [38] G. Gateshill, K. Kucharska-Pietura, J. Wattis, Attitudes towards mental disorders and emotional empathy in mental health and other healthcare professionals, *The Psychiatrist* 35 (3) (2011) 101–105.
- [39] M. Ptaszynski, Towards context aware emotional intelligence in machines: computing contextual appropriateness of affective states, in: *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 2009, July 11.
- [40] M. Jesuthan, Singapore's free AI therapy-bot is as problematic as you'd think, *Rest of World* (November 29, 2022). <https://restofworld.org/2022/free-therapy-chatbots-singapore/>.
- [41] M.M. Linehan, Validation and psychotherapy, in: A.C. Bohart, L.S. Greenberg (Eds.), *Empathy Reconsidered: New Directions in Psychotherapy*, American Psychological Association, 1997, pp. 353–392, <https://doi.org/10.1037/10226-016>.
- [42] M.J. Zielinski, J.C. Veilleux, The Perceived Invalidation of Emotion Scale (PIES): development and psychometric properties of a novel measure of current emotion invalidation, *Psychol. Assess.* 30 (11) (2018) 1454–1467, <https://doi.org/10.1037/pas0000584>.
- [43] M.M. Linehan, *Cognitive-behavioral Treatment of Borderline Personality Disorder*, Guilford Press, New York, NY, 1993.
- [44] M. Haslam, J. Arcelus, C. Farrow, C. Meyer, Attitudes towards emotional expression mediate the relationship between childhood invalidation and adult eating concern, *Eur. Eat Disord. Rev.* 20 (2012) 510–514, <https://doi.org/10.1002/erv.2198>.
- [45] D. Sells, R. Black, L. Davidson, M. Rowe, Beyond generic support: incidence and impact of invalidation in peer services for clients with severe mental illness, *Psychiatr. Serv.* 59 (2008) 1322–1327, <https://doi.org/10.1176/appi.ps.59.11.1322>.
- [46] S.J. Linton, K. Boersma, K. Vangronsveld, A. Fruzzetti, Painfully reassuring? The effects of validation on emotions and adherence in a pain test, *Eur. J. Pain* 16 (2012) 592–599, <https://doi.org/10.1016/j.ejpain.2011.07.011>.
- [47] M.B. Kool, H. van Middendorp, M.A. Lumley, J.W. Bijlsma, R. Geenen, Social support and invalidation by others contribute uniquely to the understanding of physical and mental health of patients with rheumatic diseases, *J. Health Psychol.* 18 (1) (2013) 86–95, <https://doi.org/10.1177/1359105312436438>.
- [48] T. Dunn, The dystopian cheeriness of Singapore's useless, gaslighting AI therapist chatbot, *Boing Boing*, <https://boingboing.net/2022/12/04/the-dystopian-cheeriness-of-singapores-useless-gaslighting-ai-therapist-chatbot.html>, 2022, December 4.
- [49] L.F. Barrett, R. Adolphs, S. Marsella, A.M. Martinez, S.D. Pollak, Emotional expressions reconsidered: challenges to inferring emotion from human facial movements, *Psychol. Sci. Publ. Interest* 20 (1) (2019) 1–68.
- [50] T. Watt Smith, *The enigma of emotions*, in: E.J. Ward, R. Reuvers (Eds.), *Enigmas*, Cambridge University Press, 2022.
- [51] Institute of Positive Design, *Emotion typology*, Retrieved January 5, 2023, from, <https://emotiontypology.com/>, 2022.
- [52] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: a comprehensive survey of methods and applications (arXiv:2209.00796), arXiv (2022). <http://arxiv.org/abs/2209.00796>.
- [53] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, R. Lowe, *Training Language Models to Follow Instructions with Human Feedback*, 2022 arXiv preprint arXiv:2203.02155.
- [54] E. Vedernikova, P. Kuppens, Y. Erbas, From knowledge to differentiation: increasing emotion knowledge through an intervention increases negative emotion differentiation, *Front. Psychol.* 12 (2021) 703757, <https://doi.org/10.3389/fpsyg.2021.703757>.
- [55] D. Keltner, A. Cowen, A taxonomy of positive emotions, *Current Opinion in Behavioral Sciences* 39 (2021) 216–221.
- [56] J.D. Lomas, A. Lin, S. Dikker, D. Forster, M.L. Lupetti, G. Huisman, J. Habekost, C. Beardow, P. Pandey, N. Ahmad, K. Miyapuram, T. Mullen, P. Cooper, W. van der Maden, E.S. Cross, Resonance as a design strategy for AI and social robots, *Front. Neurobot.* 16 (2022) 850489.
- [57] M.G. Brown, N. Carah, B. Robards, A. Dobson, L. Rangiah, C. De Lazzari, No targets, just vibes: tuned advertising and the algorithmic flow of social media, *Social Media + Society* 10 (1) (2024) 20563051241234691.
- [58] P. Grietzer, *A theory of vibe*, *Glass Bead* 1 (2017).
- [59] N. Benaich, I. Hogarth, *State of AI report*, London, UK, <https://www.aiunplugged.io/wp-content/uploads/2023/10/State-of-AI-Report-2023.pdf>, 2023.
- [60] M. Klenk, *Ethics of Generative AI and Manipulation: A Design-Oriented Research Agenda*, Available at: SSRN 4478397, 2023.
- [61] M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3, *Proc. Natl. Acad. Sci. USA* 120 (6) (2023) e2218523120.
- [62] T. Hagendorff, *Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods*, 2023 arXiv preprint arXiv:2303.13988.
- [63] J.E.T. Taylor, G.W. Taylor, Artificial cognition: how experimental psychology can help generate explainable artificial intelligence, *Psychonomic Bull. Rev.* 28 (2) (2021) 454–475.