AI + Creativity + Humor: What We Found?

Zhuolun Zhong

December 21, 2023

1 Background and project overview

A recent study has introduced the idea of using humor data to assess Large language model AI performance. While AI's performance falls short of humans in various tests, the results are still noteworthy. For instance, AI sometimes outperformed humans in tasks requiring the most sophisticated understanding (explaining why a joke is funny), even identifying references that human participants did not recognize (Hessel et al., 2023). This result suggests that although AI cannot surpass humans in complex mental tasks involving humor comprehension, there is a trend toward performance close to human levels rather than nonsensical outputs. Humor is a relatively understudied and mysterious topic because it is highly associated with spontaneous creativity, making it challenging to design ideal experiments involving human participants. In recent years, the rapid development of large language models has provided new perspectives and directions for researching humor.

Humor, with the laughter followed, is a universal human experience (Apte, 1987). The combined study of these aspects can reveal the nature of certain cognitive functions in humans. Graeme Ritchie, a Scottish linguist and AI researcher, suggested that AI investigations of humor can not only help to clarify theories of humor, but can also lead to important discoveries about human intelligence, language, problem-solving, and information processing more generally (Ritchie, 2001, 2004; Ritchie et al., 2007). In the past, research on humor and AI yielded little satisfactory results. However, with the advent of GPT-4, which can now pass the Turing test (Jones & Bergen, 2023), it is time to challenge the realm of humor once again using AI, aiming to unravel the uniqueness of human nature.

On the other hand, research suggests that humor plays an important and complex role at different stages of an interpersonal relationship (Caird & Martin, 2014). Making people think more humorously can bring about significant positive effects. Therefore, exploring humor's value is not confined solely to human cognition but extends to social and personal values. The challenge has always been the elusive nature of humor, making it difficult to grasp and define.

Our project is based on data from the classic humor activity, The New Yorker Cartoon Caption Contest (Jain, Jamieson, Mankoff, Nowak, & Sievert, 2020). Firstly, we aim to comprehend humor through various tests and models. It includes performance tests of artificial intelligence, feedback tests from human participants, predictions from word embedding models, predictions from neural networks, incongruity tests, visualizations of topic models, and more. Overall, these tests have failed to establish a reliable computational model theory for humor or accurately predict humor. However, our work holds potential value for future work in understanding humor. Secondly, we integrate AI to design interactive interfaces that assist participants in creating captions for The New Yorker Cartoon Caption Contest. We employed a specific prompt strategy, and the interactive experience in the tests yielded positive results. This interface can be directly applied for comprehensive humor caption creativity effectiveness testing in the future.

My role in the team is primarily as the intellectual lead. I formulate questions, design experiments and programs, conduct literature reviews, and engage in philosophical analysis. I provide background context with literature references and analyze the results in the report. More detailed information about the results should be available in the reports submitted by other team members and Github files. In the experiments conclusion section, I propose that not only does producing humor require human creativity, but also perceiving humor relies on human creativity. Across different future work sections, I provide thoughts on future experiments involving human participants.

2 Motivation and goals

My primary motivation is to enhance my research experience in cognitive psychology related to connectionism and computational models. Although I have a broad understanding of various fields in psychology, the research approach based on connectionism is new to me. Over a year, I acquired relevant statistical and machine-learning skills. I also engaged in

extensive literature reviews and research design during relevant seminars. With thorough preparation, I desire a research topic based on language statistics. Therefore, my primary goal is to proficiently and flexibly apply various language statistical analysis methods to test hypotheses about humor data.

My secondary motivation stems from a personal interest and curiosity about humor itself. Humor is a fascinating and mysterious phenomenon, and despite its supplemental value to human social life, scientifically investigating it presents sufficient challenges. On the one hand, I enjoy the challenge of exploring the unknown, and on the other hand, I also appreciate and excel at using humor to counteract the pressures of serious work. Thus, my secondary goal is to gain insights of humor through research and potentially even provide a better computational model for humor.

3 Data

The New Yorker Cartoon Caption Contest is one of the most popular features of The New Yorker magazine. The contest releases a cartoon each week, and readers are invited to submit their own captions for the cartoon with the goal of being as humorous as possible. Readers can also vote on the submitted captions for humor ratings.

Rank	Caption	Mean	Precision	Total votes	"Unfunny" votes	"Somewhat funny" votes	"Funny" votes
0	Lately, my meltdowns are coming more frequently.	1.7554	0.02467	977	439	338	200
1	My friends complain that I'm frosty and self serving.	1.7279	0.02210	1301	638	379	284
2	I scream, you scream, we all scream. It's like no one has actual conversations anymore.	1.7265	0.01862	1850	915	526	409
3	Everyone feels sorry for the kid who dropped the ice cream, but no one ever asks, "Is the ice cream OK?"	1.7133	0.01718	2229	1144	580	505

Figure 1: captions data

The data we utilized consists of humor ratings from The New Yorker Cartoon Caption Contest, which involves readers voting on submitted captions (Jain et al., 2020). The dataset has over two million captions from cartoon contest 510 to 876. Readers have three choices for a caption: funny (3 points), somewhat funny (2 points), and unfunny (1 point). The mean for a caption is calculated by dividing the total score by the total number of votes, and precision is measured by the standard deviation of the mean (Figure 1). Typically, each contest receives several thousand caption submissions, but the total votes increase rapidly over time. The contests closer to the present have more public votes than past contests.

Regarding the distribution of mean values (funniness) for all captions, only a minimal number of captions reached two or above, and many contests had no captions rated two or higher. For a single contest, there are usually dozens of captions above 1.5 (around 1 percent), hundreds between 1.5 and 1.3 (less than 10 percent), and the remaining thousands below 1.3 (more than 90 percent). Since mean values represent mass data, predicting numerical values is challenging even for humans; it becomes even more difficult for computational models and AI to do the same task. Therefore, we used 1.5 and 1.3 as thresholds to label captions as funny, somewhat funny, and unfunny. Models and AI performance in predicting humor will be evaluated based on label accuracy rather than variance.

4 Experiment 1: GPT response on humor rating

4.1 Introduction

Humor is closely related to implicit information, making it challenging to capture solely from literal expressions, such as sarcasm. Understanding humor can refer to two different scenarios. First is understanding why something is humorous and logically deducing the connection between implicit and literal information. Second is quickly perceiving implicit information and emitting laughter. The former is structured and reflective, while the latter is intuitive and reactive (Levine & Redlich, 1955). Explaining why something is funny does not necessarily elicit laughter, or at least elicit lesser laughter compared to the extent of an intuitive response. Previous research has explored understanding humor in the first scenario, explaining why something is funny (Hessel et al., 2023), but whether this constitutes a genuine understanding of a joke is debatable. Understanding humor in the second scenario is more common for humans, and not all forms of humor involve language or can be accurately interpreted (such as performance-based humor). Even for the same humorous content, not all explanations are entirely consistent in their interpretation.

The first experiment tests the current AI's performance in direct humor judgment tasks. Considering that AI's performance in the context of why funny is not as good as that of humans, this experiment assumes that AI may perform even worse in humor judgment tasks.

4.2 Method

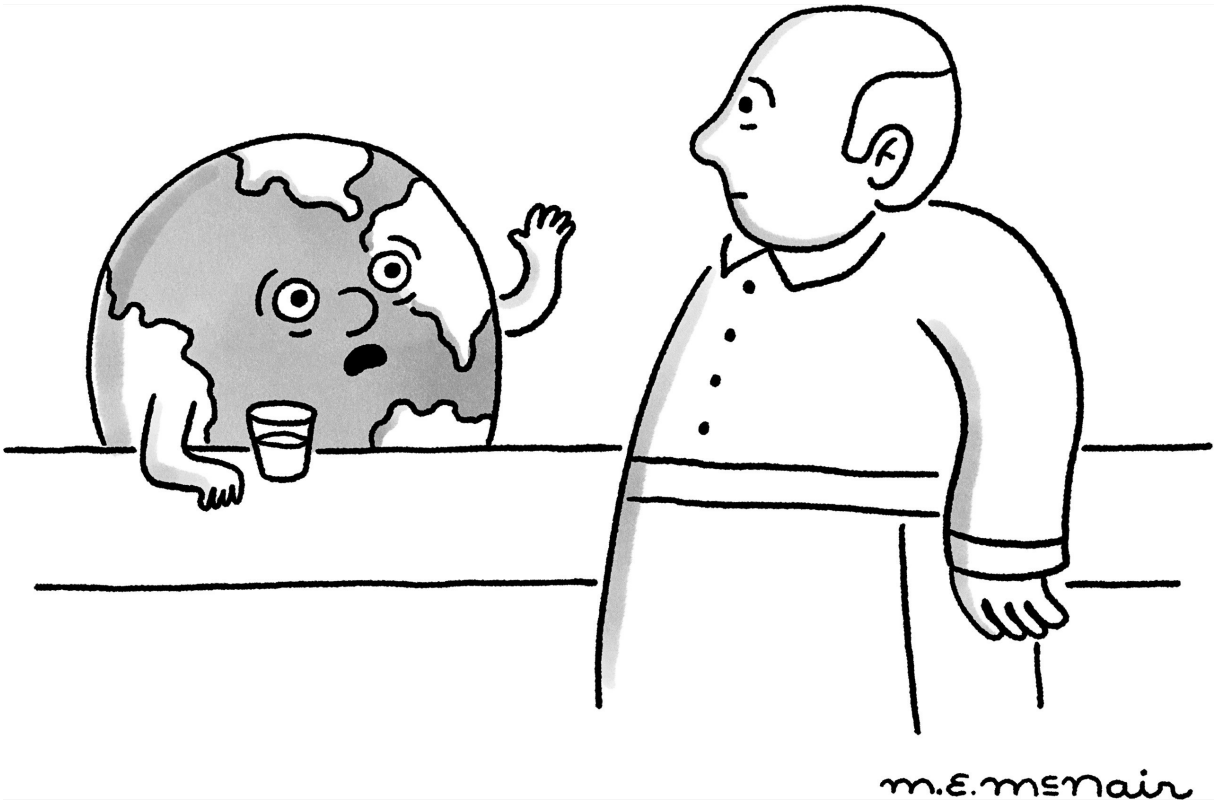

Figure 2: Cartoon contest 866

This experiment used data from contest number 866 (Figure 2). Each humor label was randomly split in a ratio of 2:8, with 20% forming the test set and 80% forming the training set. For example, ten of fifty funny captions would be placed in the test set, while the remaining forty would be placed in the training set. Although not every test would necessarily use the training set and the extent to which it was used varied, each test would employ the exact same test set.

The entire humor dataset is huge, so we chose the number 866 contest as the primary one for several reasons. Firstly, older matches may already be included in GPT's training data, so we opted for more recent ones. Secondly, computer vision was a bottleneck for model performance (Hessel et al., 2023), but today, the model's vision has made significant progress. We want to test the model's performance with original image inputs like real humans. The content of cartoon number 866 is concise, and in testing, GPT-4's descriptions perform exceptionally well, accurately covering all key elements. Third, the number 866 contest has a substantial number of rating votes (over one million), making the funniness data more reliable.

The experiment tested GPT-3.5 and GPT-4. The task involved requiring the models to make judgments based on selections from three funniness labels (funny, somewhat funny, unfunny). The textual descriptions of the cartoons were generated by the GPT-4, and human inspection confirmed the accuracy of the content.

For GPT-3.5, two approaches were taken: using the API with direct prompts and training a fine-tuned model then prompts. Several fine-tuned models are trained with different data. As for GPT-4, five approaches were taken: description input prompts, image input prompts, description input prompts with training data, image input prompts with training data, and prompts with no cartoon content. The website interface was used to upload data CSV files for prompting.

Comparing the generated results with the actual labels based on mean value will yield accuracy. The analysis will consider various scenarios to understand why accuracy might differ.

4.3 Results

GPT-3.5's direct response is always "somewhat funny". Due to the model's tendency to get stuck with repetitive prompts, the accuracy across the entire test set is inconclusive.

GPT-3.5 fine-tuning model response highly depends on the labeling ratio in training data. The different ratio combinations in the training set will be expanded and reflected in the results of the test set. For example, a ratio of 1:3:6 (funny, somewhat funny, unfunny) in the training set would lead to results in the test set of 0:0:1.

GPT-4 exhibits different outcomes under two input conditions. After repeated testing, the performance mean and standard deviation of text description input are significantly worse than those of image input. The poor performance in text description input is attributed to subjective modes, as determined by the judging method reported by GPT-4. No significant performance changes were observed in either input condition during the training condition compared to no training condition. The control group, which had no cartoon input, performed the worst. However, all GPT-4 response tests only show significant accuracy in labeling responses as unfunny, with GPT-4 rarely responses as funny or somewhat funny.

4.4 Discussion

From the performance of GPT-3.5, it appears that it has no sense of humor. Its performance is even worse than random guessing because it responds in a fixed pattern (all somewhat funny or pattern from training data), ignoring the content of the task.

The performance of GPT-4 can only indicate that it knows what is unfunny. Unlike GPT-3.5, GPT-4 doesn't limit its responses to unfunny options or get influenced solely by the training set. It attempts to respond with funny options, showing a weak level of accuracy in doing so.

In any case, AI has not demonstrated an ability to judge humor close to humans, and there is considerable controversy over whether it even approaches human performance. These results align with expectations. Firstly, research suggests that humans have specific neural foundations for humor (Yamao et al., 2015). Secondly, information involving metaphors cannot be learned directly.

The most controversial aspect lies in whether GPT-4 truly understands what is unfunny or can predict what is unfunny. Knowing what is not does not guarantee an understanding of what is. Knowing what is not a unicorn doesn't ensure an understanding of what a unicorn is. Moreover, there are differences in the learning processes between large language models and humans. Therefore, after comparing the performances of GPT-3.5 and GPT-4, it is suggested that GPT-4 has developed the ability to know what is unfunny but still has no clues about what is funny. We believe that the critical aspects present in humans, which current AI designs have failed to consider, are responsible for these results. In the subsequent conclusion sections, I will explore this missing aspect: creativity.

4.5 Future work

The progress of AI is advancing rapidly, making it worthwhile to regularly test its humor judgment capabilities. The funniness is calculated from voting data. For an individual human, the experiment of determining how much accuracy can be achieved in the same judgment task is worth exploring. In the future, recruiting participants for genuine human-AI competitive testing is considerable.

There is a distinction between the textual descriptions of images and original images as information input for humans. Recruiting participants to test the funniness means voted by humans in situations where only textual descriptions are provided would be interesting. However, before committing to visual input, AI's visual capabilities still require further exploration and testing. Skepticism should be maintained when using complex hybrid models to ensure the model's performance and output reliability.

5 Experiment 2: Human response on humor theories

5.1 Introduction

From Experiment 1, it can be concluded that large language model AI still exhibits a significant gap compared to humans in humor judgment tasks. To comprehend this disparity, Experiment 2 shifted to humans, attempting to uncover clues from

the cognitive information processes involved in humor. I delved into textbooks on humor psychology; there are three main contemporary theories of humor, each with its own explanatory power and limitations (Martin & Ford, 2018).

The first is reversal theory, which can be described as three hypotheses. (1) A person must be in a playful, paratelic motivational state. (2) The humor event must produce an increase in arousal experienced in the paratelic state as fun or excitement. (3) A person must experience a cognitive synergy in which the second interpretation of a stimulus or event involves diminishment from the first interpretation (Apter, 1982). When all three conditions are met, humor occurs.

The second is comprehension-elaboration theory, which delves into specific cognitive processes in minds. It argues that understanding humor is a three-step process: interpretation, incongruity, and reinterpretation. It proposes that the amount of humor one experiences as a result of these basic comprehension processes depends on (1) the degree to which reinterpretation of an event diminishes the importance or value of the event, (2) the type and amount of cognitive elaboration that one generates in response to the reinterpretation, and (3) the degree to which the humor event is difficult to comprehend (Wyer & Collins, 1992).

The third is the benign violation theory. The central proposition of benign violation theory is that to experience humor, one must (1) interpret a stimulus or event as a violation, (2) interpret the event as benign or harmless, and (3) hold these two interpretations simultaneously (McGraw & Warren, 2010). Essentially, a violation is anything that somehow threatens a person's view of how things should be (McGraw, Warren, & Kan, 2015).

Contemporary theories of humor remain controversial since neither is universal enough for humor. For example, empirical studies of the benign violation theory only prove that jokes meeting benign violation conditions are funnier (McGraw & Warren, 2010). Besides, the theory is vulnerable to the circular reasoning of post hoc (after-the-fact) explanations (Martin & Ford, 2018). While the comprehension-elaboration theory lacks strong empirical research, another study suggests that the easier a joke was to understand, the funnier it was rated to be (Derks, Staley, & Haselton, 1998). The reversal theory is relatively conservative; although it has support from neuroscience (Yamao et al., 2015), it lacks an in-depth exploration of cognitive processes and cannot explain the semantic understanding involved in humor (Wyer & Collins, 1992). Overall, no ideal theory can be directly followed and applied to improve AI or model performance. Therefore, the question posed in Experiment 2 is: Do people need to explicitly undergo the processes and conditions described by the theory to experience humor from The New Yorker cartoon caption contest?

5.2 Method

Experiment 2 adopted an online survey format, collecting data through Prolific. Participants were asked to watch a series of jokes created by combining cartoons with captions, and for each joke, they responded to a series of questions related to humor theory. Each participant watched a total of 5 jokes. Due to many captions in the original dataset being not funny, this experiment randomly selected captions based on humor-level labels to create a dedicated pool of jokes. This pool had an equal number of funny, somewhat funny, and unfunny captions. The jokes encountered by participants were randomly selected from this pool.

I created five questions based on humor theories. The first and second questions are related to the hypotheses of the reversal theory, corresponding to the states and arousal. The third and fourth questions relate to the hypotheses of the benign violation theory, corresponding to the violation and benign (accept). For violation, six aspects were measured: physical threat, identity threat, illogical, contrary to expectation, bad, and incorrect. The last question involves whether participants experienced two different interpretations (jump), a point emphasized by all three theories. No specific questions targeted the comprehension-elaboration theory because the processes described by the comprehension-elaboration theory are challenging to incorporate into a survey. In pre-testing before the survey was released, feedback indicated that the question created by the comprehension-elaboration theory was too complex and confusing for participants. Participants were required to evaluate the subjective funniness they experienced on a scale of 1 to 100.

After collecting the survey data, linear regression was employed to test theories.

5.3 Results

A total of 30 participants were recruited, but two had severe data missing issues and were excluded. In the end, feedback for a total of 140 jokes was obtained.

The first model for the reversal theory: lm(funny~state*arousal*jump). The only significant coefficient is arousal.

The second model for the benign violation theory: lm(funny~violation*accept*jump). The only significant coefficient is accept.

The third model for part of the comprehension-elaboration theory: lm(funny~jump). The coefficient for 'jump' is significant.

5.4 Discussion

The reversal theory did not receive confirmation in the experiment. All three predictors in the first model should have been significant, but only arousal showed significance. Furthermore, there are doubts about the significance of arousal because arousal is a physiological concept that cannot be directly measured in a questionnaire survey. As a substitute, subjective responses with a simple question were employed. The failure of the reversal theory may be attributed to criticisms of the comprehension-elaboration theory: the reversal theory lacks an assessment of funniness level (Wyer & Collins, 1992). In other words, the reversal theory did not account for differences in the funniness levels of different jokes. Therefore, if the survey's method of measuring humor is limited to choosing between funny and unfunny, the data results may vary. Overall, the shortcomings of the reversal theory, which lacks an in-depth exploration of cognitive processes, have been supported, as evidenced by the lack of significance in the data concerning humor levels.

The benign violation theory did not receive confirmation in the experiment. All three predictors in the second model should have been significant, but only accept showed significance. The survey did not directly inquire about the term "benign" but instead used "acceptable" to avoid confusion. For jokes, a bias toward "acceptable" with a high rating is predictable. The lack of significance in violation as a predictor might be because cartoons and captions in The New Yorker Cartoon Caption Contest rarely include elements of violation. As a mainstream magazine with cartoon content, caption creators may try to avoid aggressive elements. This result supports criticisms of the benign violation theory, suggesting that benign violation is an additional, unnecessary prerequisite for humor.

The comprehension-elaboration theory asks for a more complicated response on humor rating, and the "jump" predictor from the third model failed this requirement. Therefore, as criticized, the comprehension-elaboration theory lacks empirical research evidence and is challenging to investigate in experimental studies (Derks et al., 1998). The significance of the "jump" predictor in the third model suggests that the cognitive processes proposed by the comprehension-elaboration theory may exist. However, the specific cognitive processes might happen too quickly to be consciously realized.

In summary, the results of Experiment 2 support criticisms of the three contemporary humor theories. Regarding the research question, the results answer with no. Concerning the initial research motivation, the results do not offer an ideal next step. The mystery and difficulty of humor, revealed in Experiment 1, are reinforced in Experiment 2.

5.5 Future work

Based on the research question posed by Experiment 2, more detailed questions and surveys could be conducted in the future. For example, a test of the reversal theory using only funny and unfunny options. Another possibility is a content analysis-based test of the comprehension-elaboration theory, relying on descriptive responses.

Additionally, having AI attempt a similar survey could be intriguing, but the reason for doing so needs to be clarified. Overall, the most crucial aspect of Experiment 2 lies in the extensive literature review, providing an overarching understanding of existing humor research and outcomes. Humor lacks a unified theory; the so-called contemporary theories seem antiquated from a psychological perspective. In the future, attention should be focused more on conceptualizing new theories.

6 Experiment 3: Machine learning humor level prediction

6.1 Introduction

According to Ritchie, most of the existing humor theories are too vague and imprecise for computational application(Ritchie et al., 2007). This provides some support for the failure in Experiment 2. The results of Experiment 1 also failed to capture any clues to humor successfully. Therefore, the new question is: Can machine learning methods accurately predict humor solely from captions? The curiosity for Experiment 3 arises from the question of whether there is some pattern of humor in linguistic information when cognitive processes are ignored.

6.2 Method

Considering that there are too few funny captions in a single contest to provide sufficient data for model training, data from contests 850 to 870 were used. Subsequently, captions with different labels were extracted in a 1:1:1 ratio and separated into the training and test sets.

Two machine learning methods were employed: word embeddings and neural networks. Since the original data is linguistic information, the models utilized the training set directly for word embeddings. For the neural network method, a tool called LIWC-22 (Tausczik & Pennebaker, 2010) was used to transform captions into over a hundred features, and the original captions were removed. The neural network model was trained using these features.

Model parameters were adjusted, generating multiple models for both methods. The results will be evaluated based on the accuracy of each label and the overall accuracy.

6.3 Results

Despite noticeable variations in the accuracy of the three labels under certain parameter settings, the overall accuracy hovers around the level of random guessing (33%) for both methods in all models.

6.4 Discussion

Figure 3: Cartoon contest 523

From the results, it can be deduced that machine learning has failed to learn any underlying patterns from the data. At least linguistically, there is no discernible pattern of humor in captions.

Upon closer inspection of the data, it is observed that similar captions exhibit different levels of humor (Table 1). The following examples with their funniness mean are from Contest 523 (Figure 3):

Table 1: Captions with mean

captions	mean
You ordered from the wrong Amazon.	1.8912
I think I ordered from the wrong Amazon.	1.8098
we might have ordered from the wrong Amazon.	1.6794
I believe you ordered from the wrong Amazon.	1.6656
Looks like I ordered from the wrong Amazon.	1.5959
honey i think you ordered from the wrong Amazon.	1.5263

These six captions are nearly identical in content, with the remaining differences mostly consisting of some stop words (common words considered for removal in word embeddings). However, the humor level is significantly different among these six captions, and human intuition aligns with their respective rankings. For example, "I believe" is redundant and boring, "You" sets a better context than "I think". Clearly, neither word embeddings nor neural networks can capture the significant impact of these subtle differences. Because the captions in a single contest are always centered around the cartoon, there will inevitably be content similarities among captions. Considering that in Experiment 1, GPT-4's performance was the worst when there was no cartoon input, the results of Experiment 3 aligned with expectations. The cartoon content as the context plays an important role in the sense of humor.

6.5 Future work

Continuing similar explorations in machine learning in the future is not recommended.

7 Experiment 4: Incongruity test

7.1 Introduction

Back to humor theories, the experience of humor appears to be predicated on two cognitive-perceptual processes activated by characteristics of a humor stimulus and the social context in which it is encountered: (1) perception of incongruity and (2) appraisal of incongruity in a nonserious humor mindset (Martin & Ford, 2018). While the first characteristic is confirmed by all three theories and supported by the third model from experiment 2, the second characteristic becomes controversial for different theories and did not receive support from experiment 2. The reversal theory describes the humor mindset as a mental state. The comprehension-elaboration theory describes the humor mindset as a complex understanding process. The benign violation theory describes the humor mindset as a benign attitude towards violations. I argue that the humor mindset should not be treated independently for two reasons. Firstly, mindset is a concept that lacks a precise definition and cannot be easily quantitatively measured. Secondly, both the perception of incongruity and the humor mindset are subjective to humans; a subjective experience (humor mindset) based on another subjective experience (perceptive of incongruity) is redundant. Therefore, after conducting the previous experiments and reviewing the literature, I believe a fusion of the two characteristics would be worthwhile. In other words, a nuanced humor mindset may not be necessary, and perceiving incongruity in specific patterns may be sufficient to trigger a humor experience.

The discovery and navigation of possibilities involve more than the simple, separate representation of particular future or action-outcome consideration (Poulsen & DeDeo, 2023). We are often called upon to hold multiple, incompatible possibilities in mind and prepare for them in action, an ability that appears early in childhood and may well be unique to our species (Redshaw & Suddendorf, 2016). This ability is also described as representing a "matrix of maybe" (Baumeister, Maranges, & Sjåstad, 2018). Works have been dedicated to understanding how humans navigate and constrain the possible futures into this matrix (Phillips, Morris, & Cushman, 2019; Kvavilashvili & Rummel, 2020; Cole & Kvavilashvili, 2021; Sjåstad & Baumeister, 2023). When there is a noticeable disparity between reality and future matrix, incongruity experience arises. Incongruity can trigger various responses, including fear, confusion, surprise, and more, with humor being just one of the possibilities. Under the view of the future matrix, a fusion of the two humor characteristics becomes possible in a new direction. While reactions to incongruity are diverse, various responses may exhibit distinct matrix patterns, such as differences in predicting probability ranges or variations in predicted content.

Given large language models' ability to produce human-like text (Bail, 2023; Dillion, Tandon, Gu, & Gray, 2023; Binz & Schulz, 2023), they can serve as tools to check the future matrix to address incongruity patterns for humor in a given context (cartoon). The situation of different levels of humor produced by similar texts mentioned in Experiment 3 could potentially yield significant results when using large language models as tools. Experiment 4 involves using a large language model to generate token probabilities and includes three tasks: detecting specific probability ranges for punch lines in captions; comparing the changes in probabilities with and without context (description of cartoons); finding a computational model that can be used to predict humor.

7.2 Method

Due to limitations in technology, resources, and time, GPT-2 was utilized for the experiment. GPT-2 processes token IDs and generates token IDs, and it is convenient to get its prediction probabilities on all token IDs. Perplexity is calculated based on the token's actual probability in the context and GPT-2's prediction probabilities for all toke IDs in that position.

The method involves manually examining the perplexity in captions. First, the focus is on whether the perplexity generated by GPT-2 aligns with human perception. Second, the focus is on the changes in perplexity when there is no context. Third, the focus is on attempting to summarize patterns through observation.

7.3 Results

In the case of the simplest humorous captions, the perplexity generated by GPT-2 broadly aligns with human perception. However, the specific values do not accurately predict differences in humor levels, as mentioned example in Experiment 3 (Table 1). For instance, the perplexity of the punch line "Amazon" does not reflect the correct ranking of humor level, let alone consider a linear relationship.

Additionally, GPT-2 sometimes shows greater perplexity for predicates than punch lines. For humans, the predicate should be the part that sets up the expected context, generating a noticeable but not overly pronounced level of perplexity. In the case of unfunny captions, GPT-2 sometimes displays perplexity patterns similar to those seen in funny captions but entirely inconsistent with human perception. Most importantly, due to the instability in perplexity with GPT-2, observing any pattern differences between unfunny and funny captions is impossible.

A similar situation arises in context detection. Context can sometimes enhance the alignment between GPT-2 perplexity and human perception when dealing with simple cautions. However, this result is inconsistent, and the data becomes uncontrollable when facing slightly more complex captions.

The experiment did not extend to a larger dataset because the performance of GPT-2 could not help yield a computational model suitable for the hypothesis.

7.4 Discussion

Experiment 4 did not produce results that highly align with human perception, supporting the viewpoint that new ways to explore the latent spaces of human possibility need to be tempered with a certain conservatism (Poulsen & DeDeo, 2023). LLMs cannot replace ordinary psychological experiments (Dillion et al., 2023) or correct their faults (Poulsen & DeDeo, 2023). While the results of Experiment 4 can be largely attributed to the performance limitations of GPT-2, it is important to note that large language models and humans explore the space of possibilities in different ways (Poulsen & DeDeo, 2023). Caution should be exercised when using large language models to develop computational models for humor.

7.5 Future work

Considering the explanatory power of the future matrix and the partially aligned results with human perception in Experiment 4, much work still needs to be done in further researching humor from this new perspective. However, the emphasis should return to human participants. It must be emphasized that in psychology, the focus always remains on humans, and humor is a uniquely human phenomenon. Future work should remember the subject of study and the centrality of humans in understanding humor.

It's possible to revisit perplexity detection tasks using more advanced large language models in the future. However, as mentioned in the discussion section, investing excessive effort in this direction may not be necessary.

Subsequent tasks could involve similar testing on humans, having participants provide perplexity ratings word by word, followed by an overall assessment of humor and recording descriptive psychological processes.

8 Conclusion

8.1 Background

Experiments 1 and 2 showed me that solely focusing on humor or relying on AI has limitations. Therefore, I chose to take a cognitive psychology course. In the course, my emphasis has been on exploring cognitive architectures and unique human cognitive abilities to find the cognitive foundations of humor. During this process, I narrowed down the topic to creativity. In the conclusion section, I will integrate findings from four experiments to discuss the relationship between humor and creativity and potential future research paths.

8.2 The puzzle of creativity

Although the conceptual discussion of creativity should ideally fall within the realm of philosophy, creativity has yet to receive widespread attention from philosophers, even within aesthetics, the field most closely related to creativity (Gaut, 2010). On the other hand, in response to J. P. Guilford's claim in his 1950 APA presidential address (Simonton, 2000), psychology has generated substantial research on creativity. However, these efforts tend to be discrete, specialized subfield studies, resulting in a fragmented landscape and a lack of coordinated, unified knowledge construction (Glăveanu, 2014).

This situation arises because creativity is an exceptionally challenging topic. What is creativity? The most commonly used definition in psychological experiments is "the production of effective novelty" (Cropley, 1999; Lubart, 2001; Mumford, 2003; Plucker, Esping, Kaufman, & Avitia, 2015). However, defining creativity in terms of novelty and value (Weisberg, 1993) needs to be revised. Firstly, novelty is a relative concept over time; something novel at one point will not be eternal, as novelty tends to diminish over time. If novelty is adopted as the defining criterion for creativity, it would classify almost everything as creative (Hausman, 1979). Secondly, substituting originality for novelty is problematic because creative productions are not generated out of thin air but have strong connections to existing known entities, making the line between original and pre-existing content blurry (Runco, 2023). Lastly, value is a retrospective judgment, and the value of a creative production cannot be assessed before its existence in time.

I argue that the premise for discussing human creativity involves accepting indeterminism, meaning that causes do not constrain the future to a single path. It implies uncertainty and probability, which ensure the novelty. In contrast, determinism implies that humans are not creating but merely producing. Considering the uncertainty principle in quantum mechanics, indeterminism in the physical world might also be true, but I will focus on human cognitive indeterminism to avoid unnecessary discussion. Human creativity is an incremental process, introducing a certain degree of novelty based on existing entities. Therefore, I define creativity as "the production of randomness against predictive model at range from 10 percent to 30 percent". The numerical range is speculative and resolves the blurry boundary issues; falling below it suggests a lack of originality, while exceeding it implies insufficient association with existing entities. The term "randomness" emphasizes the universality of creativity, resolving the timeliness issues of value and novelty and excluding the limitation of context. The term "predictive model" emphasizes the relativity of creativity, which can be relative to subjective cognition or relative to existing entities.

The definition of creativity is controversial; hence, it is reasonable to argue that AI lacks genuine and significant creativity. As mentioned by Simon (1996), there are two assertions about computers: A simulation is no better than the assumptions built into it; A computer can do only what it is programmed to do. AI has not taken creativity into account in its cognitive architecture, so it lacks creativity.

8.3 From creativity to humor

Humor is immediate, highly context-dependent, and closely tied to human creative abilities (Martin & Ford, 2018). It is obvious that producing humor requires a high level of creativity. However, I argue that appreciating and experiencing humor also relies on humans' significant creativity. This is why humor is a unique phenomenon exclusive to humans.

Experiments 2 and 4 show that human subjective cognition plays a significant role in humor judgment. Similarly, this applies to creativity. People's judgments about creativity are based on subjective predictive models. What truly associates creativity and humor is the probability pattern behind the prediction. For creativity, too much randomness beyond predictive models is considered unrelated to existing entities. In the case of humor, excessive incongruity beyond expectations is seen

as perplexing. Both creativity and humor necessitate the subjective experience of new stimuli in a certain range of probability. Both creativity and humor also experience decay in sensitivity (novelty and funniness) over time. While experiencing creativity does not guarantee the elicitation of humor, appreciating humor does require the perception of creativity. This is why using logic to understand humor does not necessarily evoke laughter, as logic seeks confirmed and reasoned outcomes incompatible with creativity's randomness. This also explains the results of Experiments 1 and 3. AI lacks the perception of creativity, making it unable to discern funny captions and rarely make funny judgments. Humor relies on novel content; therefore, machine learning has no definite pattern to master. From an evolutionary perspective, the effects of humor may serve as a foundation for further enhancing human creativity, providing positive feedback for the release of more creativity in humans. Future experiments could measure the extent of the association between an individual's subjective sense of humor and creativity. However, great caution and attention must be exercised in experimental design, as both humor and creativity have controversial definitions, making it challenging to ensure that subjective surveys collect the desired data.

How does humor differentiate itself from other responses to incongruity and creativity in terms of the future matrix? Early studies often considered aggression crucial to humor (Martin & Ford, 2018), and the benign violation theory emerged as a response to aggression (McGraw & Warren, 2010). However, the role of aggression in humor is likely to be social context-dependent rather than essential. Imagine in a desperate context where someone suddenly makes absurdly optimistic remarks, triggering a sense of humor among those around. In this thought experiment, humor arises from a situation opposite to benign violation. Optimism is considered morally encouraged, but in a desperate context, people cannot truly accept such encouragement, leading to laughter as a response. However, humor is not solely about reversal; forms like puns do not involve extreme reversal relationships. Humor takes on various forms, so it is advisable to differentiate between different forms of humor before further dissecting the cognitive processes behind humor. I believe that important factors in controlling a sense of humor include response time and attention. For instance, in appreciating the creativity of aesthetics, humans need highly focused attention, retrieving known entities from memory and expressing a sense of creativity after a certain period. In contrast, humor occurs rapidly and only requires humans to invest a little attention in the context. These factors could be considered in future human participant tests to gain a deeper understanding.

The psychological functions of humor can be classified into three broad categories: (1) emotional and interpersonal benefits of mirth, (2) tension relief and coping, and (3) social functions in group contexts (Martin & Ford, 2018). I suggest that considering the relationship between humor and creativity, humor plays a role in enhancing human creative potential. Creativity involving higher-level cognitive abilities may not necessarily overlap; for instance, aesthetic creativity may not apply to scientific creativity. However, appreciating and producing humor are foundational creative activities, and they may have predictive value for various forms of higher-level cognitive creativity. Furthermore, I advocate that attempting a detailed predictive model for humor is futile, much like how humans cannot predict what humans will create, or similar to the fading novelty of creativity over time, theories of humor are destined to maintain a certain level of ambiguity to adapt to future developments. The evidence of humor observed from human experience is likely contingent on social environments and cannot serve as a permanent assertion for humor.

8.4 The thought experiment of twin

Twin Earth is a thought experiment meant to serve as an illustration of semantic externalism. It is proposed by philosopher Hilary Putnam in his papers "Meaning and Reference" (1973) and "The Meaning of 'Meaning'" (1975). Semantic externalism is the view that the meaning is determined, in whole or in part, by external factors to the speaker. Here is the detail of Twin Earth thought experiment: If there were a twin Earth in the universe where everything is identical to our known Earth, including its inhabitants, the only difference would be that the water on this twin Earth would have a different chemical composition from the water on our Earth. Now, if a person on Earth and his identical counterpart on the twin Earth simultaneously say "water", do they mean the same thing?

Twin Earth implies the answer is no to support the semantic externalism. Here is the counter version of the thought experiment: If there is a pair of genetically identical twins in a room, with their current physical states also being identical, and after hearing a joke, one laughs while the other doesn't, is the meaning of the joke the same for this pair of twins?

I do not intend to engage in philosophical discussions, but I would like to use this example to emphasize the significance of subjective mental activities in humor. Without considering the individuals' perspectives, the humor analysis is incomplete.

9 Humor creativity assistant interface

9.1 Introduction

The purpose of understanding and predicting humor extends beyond mere curiosity. From a value perspective, it aims to foster the development of humor creativity. We need to analyze valuable strategies for promoting humor creativity in our research on humor. Additionally, we must identify metrics from our study on humor that can be used to assess our assistant interface. I have undertaken most of the development work on the interface, including coding all essential functions. In the later stages, Lihao assisted me with code annotation, decorative adjustments, and debugging. Since Lihao has already provided a detailed description of the design aspects of the interface in his section, and given the highly research-oriented nature of this report, I will focus here on discussing the theoretical support for the effectiveness of the two main functions in fostering humor creativity.

9.2 Function 1: AI assistant

Some may argue if AI does not comprehend humor and creativity, how can it assist humans in tasks related to humor and creativity? Artificial tools do not need to possess the ability to autonomously solve problems; their mission is fulfilled by adapting to the external environment under human use, such as a paintbrush being used for painting. However, it is evident that directly asking AI for assistance is not a purposeful way of using the task of humor creativity. Because that would simply mean handing over the task to AI for autonomous completion. Therefore, what truly matters is the specific process through which creativity comes into play and the extent to which AI can assist in this process.

Creativity does not occur at the physical level; instead, it takes place at the conceptual level. Therefore, descriptions of its process are abstract and inferential. The conceptual process in which creativity occurs is often described as inspiration. In the process of literature review, I found that research suggests that inspiration stems from imitating and observing subjectively novel entities (Okada & Ishibashi, 2017). After delving into cognitive architecture and certain theories in computer science, I conceptualize inspiration as the result of parallel interactions of mental representations. Specifically, mental representations are mental imagery humans use to reconcile the internal and external environments (Simon, 1996). They take specific forms or content, portraying reality or abstraction, but are not bound by the rules of the physical world. For example, decimal representation can be a mental representation that humans use to solve mathematical problems in the external environment. However, humans can also adopt binary or hexadecimal representations in the internal environment. The physical world does not limit mathematics to the decimal system, allowing humans the freedom to flexibly and abstractly adjust their representations. Although the limitations of working memory can quickly force humans to focus on a single mental representation in a serial manner, humans can maintain multiple mental representations simultaneously in a parallel manner. Due to the adoption of different forms and content in various mental representations for different purposes, the interaction of these mental representations can trigger inspiration. For instance, reasoning logic and unicorns are entirely different mental representations. When both mental representations are given relatively balanced weight and considered simultaneously, results emerge without aligning with the predictions of either single representation.

Creativity requires inspiration, and inspiration relies on the parallel interaction of mental representations. Therefore, as a tool, AI can provide the mental representations necessary for such interactions. This is also the psychology perspective in utilizing AI: Large language models are perhaps best thought of, not as intelligences in any usual sense but as vast summaries of human behavior that take a particularly compelling form (Mitchell & Krakauer, 2023). In other words, AI can serve as an ideal repository of mental representations, offering materials for inspiration tailored to specific needs.

Our AI assistance function has established a fixed template to extract relevant information feedback from the large language model, providing users with mental representations to spark inspiration. This template primarily focuses on three design principles: strategy, randomness, and recording. Firstly, we referenced funny captions to compile dozens of strategies incorporated into the prompt for obtaining feedback from the AI. This means leveraging AI's flexible conversational abilities to convey successful humorous mental representations to the user's mind. Secondly, these strategies are randomly selected when users attempt to seek assistance from AI. This is aligned with the definition of creativity discussed in the conclusion section of the experiments. If creativity relies on randomness, it is preferable to emphasize randomness in the inspiration process. Finally, our program records captions generated by users, using labels to distinguish between those already conceived and newly created ones. Both types of records will be prompted to AI with distinction when seeking AI feedback. This approach aids AI in more quickly and accurately pinpointing relevant mental representations from the broad repository for specific scenarios.

9.3 Function 2: inspiration

The primary limitation of the AI assistance function lies in the requirement for users to provide content to receive feedback and gradually improve through continuous interaction. This implies that there may be significant challenges in the early stages of use. If a user cannot come up with any captions, the AI assistance cannot commence. Additionally, suppose a user does not find inspiration after receiving some feedback. In that case, the AI assistance may struggle to pinpoint suitable mental representations accurately, remaining in a stage of generating vague feedback. To address this issue, our interface introduces a second function to complement the shortcomings of AI assistance function.

The second function follows the same theoretical framework as the first one, so I won't go into unnecessary details. The key difference is that the second function operates independently of AI, doesn't require caption input from users, and utilizes a different template to provide inspiration. In terms of strategy, the inspiration function is simple. It selects a cartoon contest and funny captions corresponding to it from the database. This strategy aligns more closely with the conclusions drawn from inspiration research (Okada & Ishibashi, 2017). Similarly, the inspiration function incorporates randomness. Users can randomly draw cartoon contests or choose them manually. However, the three corresponding captions are randomly selected from the top 20 in the funny ranking. Despite indicating in previous experiments and discussions that appreciating humor does not require a specific mindset, specific mental representations can help create humor. These mental representations are part of one's personality and are not obtainable through direct conscious thinking. In the inspiration function, appreciating captions from cartoon contests unrelated to the task can be more beneficial for activating these underlying mental representations.

9.4 Future work

The interface has not undergone rigorous human participant experiments yet, so future testing is desired. Ideally, creating an independent cartoon contest and dividing participants into three groups—AI, humans using the interface, and humans not using the interface—could be a suitable approach. After collecting captions, classic voting methods involving other human participants could be employed to assess humor levels. However, it is important to note that individuals exhibit significant differences in humor creativity from a personality perspective, and judgments of humor levels are highly subjective. This experiment may require a substantial number of human participants to yield meaningful results.

What is more worth investigating is the developmental changes that occur in humans after long-term use of the interface. Since humor and creativity exhibit individual differences in humans, and these abilities are not static, conducting smaller-scale longitudinal tests to assess skill development over time may be better. In conclusion, there is much valuable work to be done with human participants in the future.

References

Apte, M. L. (1987). Ethnic humor versus "sense of humor" an american sociocultural dilemma. *American Behavioral Scientist*, *30*(3), 27–41.

Apter, M. J. (1982). " fawlty towers": A reversal theory analysis of a popular television comedy series. *Journal of Popular Culture*, *16*(3), 128.

Bail, C. A. (2023). Can generative ai improve social science?

Baumeister, R. F., Maranges, H. M., & Sjåstad, H. (2018). Consciousness of the future as a matrix of maybe: Pragmatic prospection and the simulation of alternative possibilities. *Psychology of Consciousness: Theory, Research, and Practice*, *5*(3), 223.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Caird, S., & Martin, R. A. (2014). Relationship-focused humor styles and relationship satisfaction in dating couples: A repeated-measures design. *Humor*, *27*(2), 227–247.

Cole, S., & Kvavilashvili, L. (2021). Spontaneous and deliberate future thinking: a dual process account. *Psychological research*, *85*, 464–479.

Cropley, A. J. (1999). Creativity and cognition: Producing effective novelty. *Roeper review*, *21*(4), 253–260.

Derks, P., Staley, R. E., & Haselton, M. G. (1998). " sense" of humor: Perception, intelligence, or expertise?

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Gaut, B. (2010). The philosophy of creativity. *Philosophy Compass*, *5*(12), 1034–1046.

Glăveanu, V. P. (2014). The psychology of creativity: A critical reading. *Creativity. Theories–Research–Applications*, *1*(1), 10–32.

Hausman, C. R. (1979). Philosophy of creativity. *Ultimate Reality and Meaning*, *2*(2), 143–162.

Hessel, J., Marasović, A., Hwang, J. D., Lee, L., Da, J., Zellers, R., ... Choi, Y. (2023). *Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest.*

Jain, L., Jamieson, K., Mankoff, R., Nowak, R., & Sievert, S. (2020). *The new yorker cartoon caption contest dataset.* `https://nextml.github.io/caption-contest-data/`.

Jones, C., & Bergen, B. (2023). Does gpt-4 pass the turing test? *arXiv preprint arXiv:2310.20216*.

Kvavilashvili, L., & Rummel, J. (2020). On the nature of everyday prospection: A review and theoretical integration of research on mind-wandering, future thinking, and prospective memory. *Review of General Psychology*, *24*(3), 210–237.

Levine, J., & Redlich, F. C. (1955). Failure to understand humor. *The Psychoanalytic Quarterly*, *24*(4), 560–572.

Lubart, T. I. (2001). Models of the creative process: Past, present and future. *Creativity research journal*, *13*(3-4), 295–308.

Martin, R. A., & Ford, T. (2018). *The psychology of humor: An integrative approach.* Academic press.

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological science*, *21*(8), 1141–1149.

McGraw, A. P., Warren, C., & Kan, C. (2015). Humorous complaining. *Journal of Consumer Research*, *41*(5), 1153–1171.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.

Mumford, M. D. (2003). Where have we been, where are we going? taking stock in creativity research. *Creativity research journal*, *15*(2-3), 107–120.

Okada, T., & Ishibashi, K. (2017). Imitation, inspiration, and creation: Cognitive process of creative drawing by copying others' artworks. *Cognitive science*, *41*(7), 1804–1837.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, *23*(12), 1026–1040.

Plucker, J. A., Esping, A., Kaufman, J. C., & Avitia, M. J. (2015). Creativity and intelligence. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.), *Handbook of intelligence: Evolutionary theory, historical perspective, and current concepts* (pp. 283–291). New York, NY: Springer New York. doi: 10.1007/978-1-4939-1562-$0_1$9

Poulsen, V. M., & DeDeo, S. (2023). *Large language models in the labyrinth: Possibility spaces and moral constraints* (Vol. 1) (No. 4). SAGE Publications Sage UK: London, England.

Putnam, H. (1973). Meaning and reference. *The journal of philosophy*, *70*(19), 699–711.

Putnam, H. (1975). The meaning of" meaning".

Redshaw, J., & Suddendorf, T. (2016). Children's and apes' preparatory responses to two mutually exclusive possibilities. *Current Biology*, *26*(13), 1758–1762.

Ritchie, G. (2001). Current directions in computational humour. *Artificial Intelligence Review*, *16*(2), 119–135. doi: 10.1023/A:1011610210506

Ritchie, G. (2004).
The linguistic analysis of jokes. doi: 10.4324/9780203406953

Ritchie, G., Manurung, R., Pain, H., Waller, A., Black, R., & O'Mara, D. (2007, December 1). A practical application of computational humour. In (pp. 91–98). (4th International Joint Workshop on Computational Creativity, IJWCC 2007 ; Conference date: 17-06-2007 Through 19-06-2007)

Runco, M. A. (2023). *Creativity: Research, development, and practice.* Academic Press.

Simon, H. A. (1996). *The sciences of the artificial.* MIT press.

Simonton, D. K. (2000). Creativity: Cognitive, personal, developmental, and social aspects. *American psychologist*, *55*(1), 151.

Sjåstad, H., & Baumeister, R. F. (2023). Fast optimism, slow realism? causal evidence for a two-step model of future thinking. *Cognition*, *236*, 105447.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Weisberg, R. W. (1993). Creativity: Beyond the myth of genius.

Wyer, R. S., & Collins, J. E. (1992). A theory of humor elicitation. *Psychological review*, *99*(4), 663.

Yamao, Y., Matsumoto, R., Kunieda, T., Shibata, S., Shimotake, A., Kikuchi, T., ... others (2015). Neural correlates of mirth and laughter: a direct electrical cortical stimulation study. *Cortex*, *66*, 134–140.