PSYCH 1026 - Psychology of Language
Research Paper

Fall 2022
Isha Puri

No Laughing Matter:
A Proposal for the Analysis of Human Reactions to AI-Generated Communication

---

Introduction and Background

Humans have, historically, never had a particularly easy relationship with transformational advances in technology. Take the printing press, for example. Or the microscope, lightbulb, telegraph, radio, or computer (Bell, 2010). In every major industrial revolution, "technophobic" humans have fought back against an uprooting of "the way things should be", displaying a certain mistrust towards major advances in technology.

The recent AI revolution is no different. AI has completely revolutionized the world we live in, and its applications have integrated themselves into our social media networks, health diagnostic practices, and even the criminal justice systems. And while realistic text generation algorithms were once considered almost impossible to build, the elusive jewel of the AI crown, recent advances in machine learning have expanded the capacities of "Natural Language Processing". AI-text generators produced by startup OpenAI like GPT2 and GPT3 (among other language generators) have developed the capability to produce communications across a wide variety of domains in a way that adheres to grammatical and semantic rules and that shockingly reaches human levels of intelligibility (Brown et al., 2020).

It is important to note that communication is simply defined as the act of imparting or exchanging information of any kind (Dictionary.com, 2022). It is not necessary that this communication must be between two humans - as long as information or sentiment is being exchanged in a given interaction, an interaction between a human and a computer can also be considered communication.

Now, these AI-text generators have tremendous use case capability in the real world. GPT3 can generate news articles, summarize papers, translate between languages, produce its own code given a question, write an original novel, and even generate jokes (Mudumba, 2022). As the scale and power of these models grow, therefore, their implications and effects on society will continue to explode. Thus, it is crucial that we study how humans respond and interact to communications delivered by such models.

Interestingly, as the power of GPT3-style models continues to evolve, studies have shown that it is not always straightforward for people to differentiate between communications produced by a human and

communications produced by an algorithm. In one 2021 study, researchers presented a gender-balanced set of 30 participants with a random selection of human and machine generated news articles (Tewari et al., 2021). Each machine generated article was around 400 words in length and was a result of the exact same GPT2 model settings. After reading each of 12 articles, the participants were asked to summarize the article and state the following: how credible they thought the article was, how engaging they thought the article was, what they believed the political leaning of the article was, and whether they thought the article was written by a human or a computer. In their results, the human-generated articles were correctly identified to have been written by a human 70% of the time, while machine-generated articles were correctly identified to be written by a machine only 53% of the time. And while these results suggested that the participants in the study could more confidently identify human generated articles than they could machine generated articles, there were a few flaws to the study that might taint the results. Because they were only given two authorship options to choose between, the participants in the study were aware that around half of the articles they were presented were written by a machine, which may have biased their view slightly. The small sample size also makes it difficult to assume these findings generalize over a large population.

Additionally, the 2021 Tewari study showed that participants assigned higher credibility scores to articles they perceived to be algorithmically-generated. These results are supported by the results of a larger 2017 study measuring participant trust in human-algorithm written articles (Jung et al., 2017). The study took a two (article authorship: human or algorithm) by two (displayed authorship: human or algorithm) approach and concluded that participants placed *higher* trust in articles written by an algorithm than those written by human journalists. The scope of this study was larger than Tewari, et al., tracking 201 people (103 male and 98 female) recruited from a national online panel with median age 39.6.

Both Tewari et al. and Jung et al. provided participants with news articles that typically covered sports, finance, and the weather. These genres were chosen because given their objective nature, it was easier to algorithmically generate such news stories. A side effect of this genre choice, however, is that the studies didn't truly test the limits of algorithmically generated news and didn't truly challenge human perception because it was the work produced by an algorithm and a human was so similar. Additionally, human perception of algorithmically-written pieces might have improved in these studies specifically because of the objective nature of the genres they were operating with - participants might have trusted algorithms to a greater extent with objective tasks such as information retrieval. This begs the question - does the human preference for algorithmically-generated content extend to all genders? Or does the genre of the content affect their perception?

The 2021 study *Artificial Intelligence versus Maya Angelou: Experimental Evidence that people cannot differentiate AI-generated from human-written poetry* attempts to answer this question for a more creative genre - poetry. The study's results solidified our proposed difference between human perceptions of algorithmically generated poetry and generated news articles. Researchers at the University of Amsterdam and Max Planck Institutes analyzed the reactions of 192 participants to various GPT2 or human-generated poems, measuring their preference for poetry and their accuracy in identifying algorithmically generated poetry (Köbis & Mossink, 2021). The 192 participants had a mean age of 29.06, and 39.1% of the participants were female. There was a continued failure by the participants to reliably identify the authorship of the poetry they were reading - overall, the participants selected the correct author 50.21% of the time, which is equivalent to random chance. Additionally, when scoring how much they liked the poems, the participants were found to slightly prefer human-written poetry to algorithmically-generated poetry, even when they

couldn't identify the source of the poem. Contrary to the author's hypothesis, however, when informed of the true authorship of the poems, the participants' aversion to algorithmically-generated poetry did *not* increase from the baseline scores they gave when they were unaware of authorship.

In general, I will study the similarities and differences between human perceptions of and interactions with communication that is generated algorithmically versus communication that is produced by another human. More specifically, how does the medium (i.e. content - newspaper article, novel, poetry, jokes, article summaries, etc) and context (i.e. being aware of authorship) affect whether or not humans are more likely to trust and have positive responses to AI-generated communication? We can see in the background studies mentioned above that the genre of communication strongly affects how humans perceive algorithmically -generated content, and I aim to study the human perception of this content in a genre that is emotionally charged.

Hypothesis and Experimental Logic

An area of creative communication whose success entirely depends on emotion is that of comedy. Dr. Tim Miles from the University of Surrey established a link between humor and emotion, stating that emotion plays "a large part in how well an audience connects and reacts to a comedian, and vice versa" (Miles, 2014). By analyzing dozens of empirical questionnaires and qualitative interviews with audience members and comedians alike (including Russell Brand and Robin Williams), he found that a live stand up show fulfills a need for feelings of truth, trust, empathy and intimacy between people". Because stand-up comedy is a performance, he states, "emotional experiences like identification, interaction, empathy, mutual therapy, well-being and a need for recognition all play an important part".

Despite this deep reliance on emotion, comedy itself is something that often follows a very successful formula. Joe Toplyn, a Harvard engineering undergrad and MBA who went on to pivot and become a comedy writer (eventually becoming head writer for both *The Late Show With David Letterman* and *The Tonight Show With Jay Leno*), has stated that the most successful late-night jokes he wrote followed easy-to-follow formulas that substituted certain references (people, places, things, etc) to form a punch line (Toplyn, 2021). Falling back on his engineering background, he created an algorithm, Witscipt, that was trained on a data set of late night monologue jokes, detecting keywords in a given text and creating a punchline. It takes as input user text and automatically generates contextual relevant jokes. The ability to automatically generate jokes is also possessed by more recent AI algorithms, including GPT2 and GPT3 (Floridi, 2020).

The area of comedy is a fascinating place to study human perceptions on communicating with computers. Despite its heavy reliance on emotion during live standup performance (Woodruff, 1997), a joke is a joke in any form, and it would be interesting to see how humans perceive jokes written by machines versus humans.

In this study, I am to analyze the following:
1) Will participants be able to identify the authorship of algorithmically-generated jokes as non-human?
2) If unaware of the true authorship of the jokes they are viewing, do participants show a preference or aversion to algorithmically-generated jokes?
3) Does perceived authorship of the joke change its perception in a meaningful way from the results of question 2?

Comedy as a form of communication is admittedly massively diverse (Masterclass, 2021), and even when limited to written jokes, there exist several types of jokes that one could use to conduct this study (Attardo & Chabanne, 1992). In this study, we will deal only with written jokes (as opposed to a live, stand-up comedy performance). In order to cover both short punchline jokes (i.e. “What’s the best thing about Switzerland?” “I don’t know, but the flag is a big plus.”) and slightly longer jokes with a more elaborate setup, we will include both in the dataset being used in the study while ensuring that the ratio of long to short jokes is balanced for both human-written and algorithmically-generated content.

My hypothesis is that participants will not be able to consistently and accurately identify the authorship of algorithmically-generated jokes. Additionally, I hypothesize that although the ratings for algorithmically-generated jokes will be slightly lower than human-written jokes when the participants are not informed of true authorship, this disparity will *not* increase when the treatment is applied and the participants are made aware of authorship.

As technology has evolved, human ability to differentiate between algorithmically generated content and human generated content - otherwise called the Turing Test (Turing, 1950) - has decreased significantly. A majority of factual/narrative content produced by GPT2 and GPT3 passes the Turing Test (Elkins & Chun, 2020), which is why I hypothesize that participants will not be able to consistently and accurately identify the authorship of algorithmically-generated jokes.

My hypothesis that the disparity between algorithmically-generated and human-written joke ratings will not increase when true authorship is revealed is partly based on the Köbis & Mossink 2021 poetry paper. Although one’s instinct is to assume that humans would strongly prefer human-written content to algorithmically-generated content, in a field as fundamentally based on emotional communication as poetry, the authors didn’t find a massive drop in ratings of algorithmically generated content when the author was revealed, and I predict that similar phenomena will occur in the genre of comedy.

Methods

My proposed experimental design would study a few related research questions. We would start with a group of around 300 participants, balanced for both gender and age (as younger participants might be more comfortable with technology). We would need a large sample size to overcome any personal biases with regards to humor, as comedy can be a very personally subjective form of communication (Ziv, 1979). They would complete a pre-questionnaire with questions that gave us information on (among other things, as to not tip them off to the exact nature of the study) their race, age, familiarity with technology, preferences for humor, etc. We would produce a large set of both human generated jokes from late night television monologues and AI-generated jokes from Witscript, or another joke-generating algorithm (perhaps GPT3). Human and computer generated jokes would be paired by topic, as to not allow the topic to bias a response.

In order to answer my three main questions, I will design my experiments as follows.

In the first experiment, half of the participants would hear the human written joke, while the other half would hear the AI generated joke (both narrated by humans). They would then note down their rating of the joke on a scale of 1 to 10. They would then be asked to guess who the joke was written by from a list of

options - pro-comedian, amateur comedian, not a comedian, AI - this list attempts to make the task of detecting AI less obvious.

This experiment will allow us to answer both questions 1 and 2. Because participants are first hearing a joke (without any context of authorship) and then both rating the jokes on a standard scale *and* guessing the authorship of the joke, we will be able to measure any statistically significant patterns in their underlying preference for algorithmically-generated or human-written jokes. We will also be able to measure how often participants can guess the correct authorship of a joke. Interestingly, we can also use this experiment to measure how a participant's *perceived beliefs on joke authorship* affect how they rate a joke. It would be significant, for instance, if a participant consistently gives jokes they believe to be algorithmically generated higher ratings than jokes they believe to be human-written, even if they are incorrect about the source of the authorship.

In the second experiment, we will solidify our answer to this last question by building a two (true authorship: algorithm and human) by two (publicized authorship: algorithm and human) experiment. As in Experiment 1, participants will be read jokes and asked to rate them on a standard scale. The experiment would be run on new participants with the same jokes, for uniformity.

		True Authorship of Joke	
		Written by Algorithm	Written by Human
Publicized Authorship of Joke	Told: Written by Algorithm	1	2
	Told: Written by Human	3	4

As one can see in the table above, this results in 4 main cases. In case 1, the joke is actually algorithmically -generated, and participants are told as such. In case 2, the joke is written by a human, but the participants are told that it is algorithmically-generated. In case 3, the joke is algorithmically-generated, but participants are told that it was written by a human. In case 4, the joke is written by a human, and participants are told as such. By analyzing the differences in average participant joke ratings between each of the cases, we can get a lot of valuable information. By comparing cases 1 and 2, as well as cases 3 and 4 (in essence comparing cases where the publicized authorship is constant but the true authorship is altered), we can see whether or not participants have innate preferences for jokes written by algorithms or humans. The results of this comparison will provide backing and support to some of the results from Experiment 1. By comparing cases 1 and 3, as well as cases 2 and 4 (in essence comparing cases where the true authorship is constant but the publicized authorship is altered), we can see how a participant's perceived beliefs on joke authorship affect how they rate a joke. Again, this will support and validate some results from the second part of Experiment 1.

By analyzing these results on a very large sample size, hopefully we will be able to get one step closer to understanding whether or not people perceive communication differently based on its source, and how that perception can change in different types of mediums.

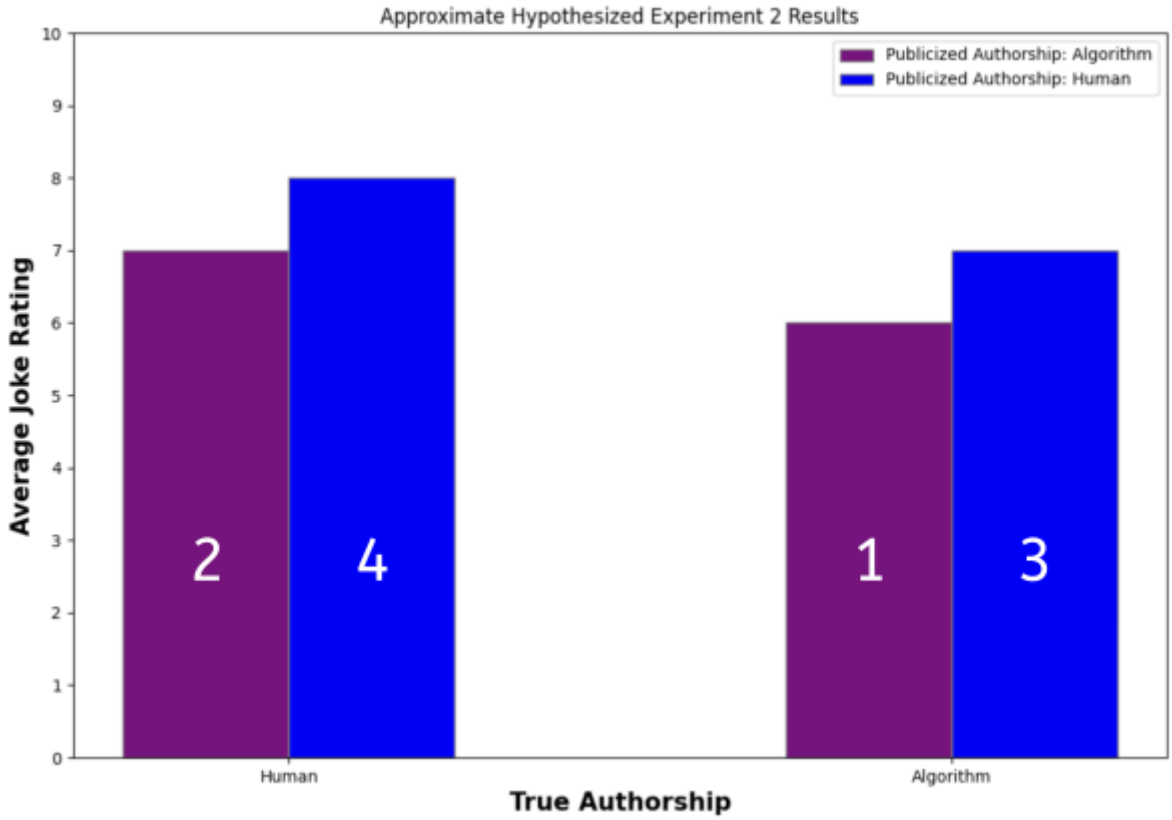
Implications of Hypothesized Results and Discussion

In the first part of Experiment 1, we asked participants to guess the authorship of a joke between the choices of "pro-comedian", "amateur comedian", "not a comedian", and "AI". When analyzing our results, we will add all guesses for "pro-comedian", "amateur comedian", and "not a comedian" into a large "human" category. I hypothesize that the overall participant accuracy in correctly guessing the authorship between human and algorithm will be around 50%, meaning that participants are *not* able to differentiate between the two. This would suggest that GPT2 and GPT3 have passed the Turing Test when it comes to comedic material, and, in a global capacity, has implications both ethically and with regards to communication. As language generation algorithms infiltrate more of our daily lives, if they are able to go undistinguished from human-generated content, we must perform large amounts of research and use abundant caution when analyzing how they impact our society and the way we communicate with each other. Additionally, if algorithmically-generated jokes are receiving similar ratings to human-written ones, we have to question what it means to be creative in communication.

Some alternative results to this part of the experiment could be as follows: 1) participants have a high accuracy in correctly identifying the authorship of *both* algorithmically and human generated jokes (which would imply that the algorithms have not yet passed the Turing Test and even though they are technologically novel, at this moment in time they will not deeply affect the way we communicate), 2) participants have a high accuracy in correctly identifying the authorship of algorithmically generated jokes but *not* human written jokes (which means participants classify most jokes as being written by algorithms), and 3) participants have a high accuracy in correctly identifying the authorship of human written jokes but *not* algorithmically generated jokes (which means participants classify most jokes as being written by humans).

In the second part of Experiment 1, we measured participants' rating of jokes on a standard scale. My hypothesized result is that there will be a slight, but not overly large, overall preference towards jokes written by humans. This would imply that perhaps, in a form of communication as hit-or-miss as humor (Rosenbusch et al., 2022), people are more picky and technology hasn't yet reached the potential of being able to completely replicate the abilities of humans. This also suggests that when we communicate, humans have high standards for humor. The completely subjective nature of humor means that everyone will find different things funny, but in general, it is still very difficult to make someone laugh (Attardo & Raskin, 1991; Bogdan, 2014). This is an important finding in the study of communication because it implies that unlike in several other genres and mediums of communication, humor can fail in its communicative goals a majority of the time. A 1986 article suggested that all successful forms of humor communication involves some sort of "buy in" on the part of the listener - a form of investment in what the speaker is saying (i.e. listening, investing in, and comprehending the "setup" of the joke in order to truly understand and appreciate the punchline". These findings would suggest that humans are more skilled at writing jokes that inherently deliver that sense of "buy in" than algorithms can (at this present point in time). This speaks to the "outgoing" part of communication - how humans write and communicate with others - instead of perception.

My hypothesized results for Experiment 2 are shown in the plot below. The case numbers from the table above are included here for consistency. As we can see, the average scores for jokes actually authored by humans are just slightly higher than the average scores for jokes actually authored by algorithms. There is not a significant difference, however, between the ratings of jokes publicized to be written by humans versus those publicized to be written by algorithms.



These findings would support the fact that participants inherently prefer jokes written by humans, which we found above in the results of part 2 of Experiment 1. (Note that this is different from passing the Turing Test, which simply asks participants to identify whether content is produced by a human or algorithm. It is possible that participants couldn't explicitly identify a joke as written by an algorithm but still have a slight preference ratings-wise for human-written jokes).

There is also not a large drop in ratings when participants find out that a joke is written by an algorithm instead of a human. From this, we could conclude that people value the content of a joke more than who wrote it - or, in simpler words, "a joke is a joke". This would be a very notable conclusion in the study of communication: in short-form punchline-based comedy, people lend more importance to the content of the communication instead of the source of the communication. Based on the aforementioned concept of emotional "buy in", these findings would suggest that short-form comedy requires less of an emotional connection between people in order to be successful in communication, and this principle could be applied to a whole host of situations. In a business context, for example, or when meeting new people, this implication suggests that telling a short-form joke might instantly make you more likable and improve the efficiency of future communication - very little emotional connection is required beforehand to make the

joke successful, and it is a good way to set up a good rapport with those you aim to communicate with. Additionally, while algorithms might not be able to replicate the high quality of human-written humor yet, when they do reach that level, these results would imply that people would be comfortable hearing jokes written by the algorithms just as long as they are as good, despite the authorship. This could lead to late night monologue jokes being written by algorithms, features on dating apps and messaging platforms to automatically come up with a good joke given some context, and even mental health platforms that use humor to diffuse complicated situations.

A 2019 study found that most people have a greater aversion to algorithms carrying out "emotional" tasks than they do to algorithms carrying out "mechanical" tasks (Castelo et al., 2019), and this hypothesized result would suggest that people view short-form "punchline-based" jokes as less emotionally intensive as other forms of comedy and/or communication. This helps contribute to our understanding of how people interact with various types of comedy, and helps solidify the notion that humor is one of the most diverse and fickle forms of communication.

To my knowledge, this is the first study to study human perception of algorithmically generated comedy content. It is crucial to understand how humans communicate, interact with, and perceive algorithmically-generated content as it becomes more commonplace in our society. Working to understand this human-computer communication dynamic will be important to develop more efficient policies to ensure the positive impact of AI. We have seen in the last decade or so just how deeply algorithms can rock the core of how humans communicate with one another (via social media, etc.), and studying the way people interact with algorithmically-generated humor is just one step in the ongoing effort to understand what factors contribute to aversion to algorithmic decision-making and generation across various domains.

Future Work

In the future, I'd like to conduct studies that take into account more demographic information about participants, and analyze how the backgrounds of participants affect their aversion to algorithmically generated content of different genres. Factors including socioeconomic status (Scherer & Siddiq, 2019), cultural background (Scott, 2019) levels of education (Yu, 2021; Riddell & Song, 2012), rural or urban residency (Haggstrom et al., 2019), and age (Kennedy & Funk, 2022) all have strong effects on how one interacts with technology, and so it would be interesting to analyze how members of these different demographics perform in the above experiments. More generally, it would be interesting to analyze how one's personal background leads one to be more or less trusting and open to new technologies, and specifically how it leads one to be receptive to algorithmically-generated content in genres that are commonly thought to require human creativity, including comedy.

Additionally, I'd like to extend this study to include standup comedy, which scientists agree is a more emotionally-charged form of comedy than short-form written comedy. Replicating the above experiments but having participants watch standup clips instead of listening to a joke would allow us to understand how humans communicate with humor in an in-person, vocal context. The setup for this experiment would be significantly more involved, however, as the researchers will have to record hundreds of standup clips.

References

Attardo, S., & Chabanne, J. (1992). Jokes as a text type. *Humor - International Journal of Humor Research*, *5*(1-2). https://doi.org/10.1515/humr.1992.5.1-2.165

Attardo, S., & Raskin, V. (1991). Script theory revis(it)ed: Joke similarity and joke representation model. *Humor - International Journal of Humor Research*, *4*(3-4). https://doi.org/10.1515/humr.1991.4.3-4.293

Bell, V. (2010, February 15). *A history of media technology scares, from the printing press to Facebook*. Slate Magazine. https://slate.com/technology/2010/02/a-history-of-media-technology-scares-from-the-printing-press-to-facebook.html

Bogdan, S. (2014). Failed Humour And Its Effects in Conversation: A Case Study. *Speech and Context*, *6*(1), 37-47.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert_voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809-825. https://doi.org/10.1177/0022243719851788

Clerwall, C. (2014). Enter the robot journalist. *Journalism Practice*, *8*(5), 519-531. https://doi.org/10.1080/17512786.2014.883116

Communication. (2022). In *Thesaurus.com*. https://Www.dictionary.com

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311-313. https://doi.org/10.1038/538311a

Elkins, K., & Chun, J. (2020). Can GPT-3 pass a writer's Turing test? *Journal of Cultural Analytics*, *5*(2). https://doi.org/10.22148/001c.17212

Floridi, L. (2020). GPT‑3: Its nature, scope, limits, and consequences. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3827044

Haggstrom, D. A., Lee, J. L., Dickinson, S. L., Kianersi, S., Roberts, J. L., Teal, E., Baker, L. B., & Rawl, S. M. (2019). Rural and Urban Differences in the Adoption of New Health Information and Medical Technologies. *Journal of Rural Health*, *35*(2), 144-154.

Jung, J., Song, H., Kim, Y., Im, H., & Oh, S. (2017). Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in Human Behavior*, *71*, 291-298. https://doi.org/10.1016/j.chb.2017.02.022

Kennedy, B., & Funk, C. (2022). Early technology adopters. *Pew Research Center*.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, 106553. https://doi.org/10.1016/j.chb.2020.106553

Lintott, S. (2017). Why (Not) philosophy of stand-up comedy? *Aesthetics*, 362-366. https://doi.org/10.4324/9781315303673-73

MasterClass. (2021, October 26). *13 types of comedy: Popular types of comedic performance*. https://www.masterclass.com/articles/types-of-comedy

Miles, T. (2019). No greater foe? Rethinking emotion and humour, with particular attention to the relationship between audience members and stand-up comedians (5:1). *The Routledge Comedy Studies Reader*, *5*(1), 115-123. https://doi.org/10.4324/9780429057526-12

Mudumba, R. (2022, July 15). *OpenAI's GPT-3 use cases in 2023! To what extent will it evolve…*. LinkedIn. https://www.linkedin.com/pulse/openais-gpt-3-use-cases-2023-what-extent-evolve-rajeev-mudumba/?trk=pulse-article

Riddell, W. C., & Song, X. (2012). The Role of Education in Technology Use and Adoption: Evidence from the Canadian Workplace and Employee Survey. *IZA Discussion Paper Series*, (6377).

Rosenbusch, H., Evans, A. M., & Zeelenberg, M. (2022). The relative importance of joke and audience characteristics in eliciting amusement. *Psychological Science*, *33*(9), 1386-1394. https://doi.org/10.1177/09567976221098595

Scherer, R., & Siddiq, F. (2019). The relation between students' socioeconomic status and ICT literacy: Findings from a meta-analysis. *Computers & Education*, *138*, 13-32.

Scott, S. (2019, October 24). *The mystery of what makes a joke funny – but only to some people*. The Conversation. https://theconversation.com/the-mystery-of-what-makes-a-joke-funny-but-only-to-some-people-125769

Tewari, S., Zabounidis, R., Kothari, A., Bailey, R., & Alm, C. O. (2021). Perceptions of human and machine-generated articles. *Digital Threats: Research and Practice*, *2*(2), 1-16. https://doi.org/10.1145/3428158

Toplyn, J. (2021). Witscript: A System for Generating Improvised Jokes in a Conversation. *Proceedings of the 12th International Conference on Computational Creativity*, 22-31.

Turing, A. M. (1950). Turing test. *PsycTESTS Dataset*. https://doi.org/10.1037/t78475-000

Woodruff, P. (1997). The paradox of comedy. *Philosophical Topics*, *25*(1), 319-335. https://doi.org/10.5840/philtopics199725111

Yu, Z. (2021). The effects of gender, educational level, and personality on online learning outcomes during the COVID-19 pandemic. *International Journal of Educational Technology in Higher Education*, *18*(14).

Ziv, A. (1979). Sociometry of humor: Objectifying the subjective. *Perceptual and Motor Skills*, *49*(1), 97-98. https://doi.org/10.2466/pms.1979.49.1.97