



OPEN ACCESS

EDITED BY

Michel Fathi,
University of North Texas, United States

REVIEWED BY

Sheena Iyengar,
Columbia University, United States
Edwin Ramirez-Asis,
Lord of Sipan University, Peru

*CORRESPONDENCE

Mark Du
✉ arsl4000@gmail.com

RECEIVED 18 April 2023

ACCEPTED 02 August 2023

PUBLISHED 23 August 2023

CITATION

Du M (2023) Machine vs. human, who makes a better judgment on innovation? Take GPT-4 for example. *Front. Artif. Intell.* 6:1206516. doi: 10.3389/frai.2023.1206516

COPYRIGHT

© 2023 Du. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Machine vs. human, who makes a better judgment on innovation? Take GPT-4 for example

Mark Du*

Department of Computer Science, National Taiwan University, New Taipei, Taiwan

Introduction: Human decision-making is a complex process that is often influenced by various external and internal factors. One such factor is noise, random, and irrelevant influences that can skew outcomes.

Methods: This essay uses the CAT test and computer simulations to measure creativity.

Results: Evidence indicates that humans are intrinsically prone to noise, leading to inconsistent and, at times, inaccurate decisions. In contrast, simple rules demonstrate a higher level of accuracy and consistency, while artificial intelligence demonstrates an even higher capability to process vast data and employ logical algorithms.

Discussion: The potential of AI, particularly its intuitive capabilities, might be surpassing human intuition in specific decision-making scenarios. This raises crucial questions about the future roles of humans and machines in decision-making spheres, especially in domains where precision is paramount.

KEYWORDS

noise, creativity, artificial intelligence, large language models (LLM), GPT-4

Introduction

Human decision-making is a complex process that involves a multitude of factors, including emotions, biases, and external influences. However, even when individuals strive to make rational decisions, they may still be prone to noise (Kahneman et al., 2021), which refers to irrelevant and random factors that can influence decision-making. This can have significant consequences in various fields, such as finance, medicine, and hiring, where noise can lead to suboptimal outcomes and undermine the effectiveness of decision-making processes. In contrast, artificial intelligence (AI) has the potential to make less noise than humans due to its ability to process vast amounts of data and apply logical algorithms to make decisions. There are some studies using GPT to examine the human mind, including semantic association (Digutsch and Kosinski, 2023), Theory of Mind (Kosinski, 2023), personality (Cao and Kosinski, 2023), decision making (Hagendorff et al., 2022), and even stock market (Lopez-Lira and Tang, 2023). This essay explores the concept of noise in human decision-making and examines how AI can help reduce noise and improve outcomes. By understanding the limitations of human decision and the benefits of AI, we can identify opportunities to enhance judgment processes and promote more effective and efficient outcomes in various domains.

In the book “Noise,” Daniel Kahneman argues that noise is a psychological factor that we don’t want. On the other hand, creativity is something that we wish to do, which also involves a process of thinking in a different aspect. There’s a lot of research focused on how to judge a new innovative idea (Grant, 2017). Creativity requires divergent thinking, which is a similar

psychological concept to noise. When it comes to judging innovative ideas, it seems that people get both bias and noisy. According to “Noise,” Kahneman argues that the AI system is noiseless. But after GPT-4 (2023) came out, things get a little bit different. Chat GPT is an AI model using a probability algorithm, which means it could generate randomness, and of course, noise. However, this article investigates that even a noisy AI system is less noisy than human judgment.

In this article, we want to compare machine prediction and human judgment, there is already some research comparing both (Kleinberg et al., 2018). Even a simple model can outperform humans (Meehl, 1954). We also compare GPT-3, GPT-3.5, and GPT-4. GPT is constantly refreshing its algorithm, past research suggests that GPT will make some judgment mistakes like a human, but a new version of GPT shows that some of these mistakes have been corrected.

Machine judgment: overcoming intuition

Noise is composed of lots of types of bias, one example is intuition. Humans make mistakes using intuition and create bias. Intuition bias refers to the tendency to rely too heavily on intuition or “gut feelings” when making decisions or judgments, rather than relying on objective evidence or systematic analysis. While some researchers suggest that gut feeling might be helpful when judging start-up projects (Huang, 2020), most behavioral economic research suggests the opposite. If the situation does not include emotion or social cues, gut feeling is not the best option to make decisions. We discover that GPT will not make the same mistake as humans.

One method to measure intuition is the cognitive reflection test (CRT; Frederick, 2005). A classic question is to ask how much a ball costs:

Prompt 1.1

A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost?

The correct answer is 5 cents, and the intuitive answer is 10 cents. Here is an example of how GPT-3.5 answer this question:

Let's assume that the cost of the ball is x .

According to the problem, the cost of the bat is \$1 more than the ball,

So: Bat cost = $x + \$1$.

The total cost of the bat and ball together is \$1.10:

Total cost = Bat cost + Ball cost $\$1.10 = (x + \$1) + x$ Simplifying the equation:

$\$1.10 = 2x + \1 $\$0.10 = 2x$ $x = \$0.05$.

Therefore, the cost of the ball is \$0.05.

So the correct answer is \$0.05.

As you can see, GPT-3.5 list the math formula to solve the question, instead of thinking about it intuitively.

Another problem is called Hospital Problem, we include GPT-3.5 and GPT-4/s answers in Appendix B, and they both answer correctly. GPT-3.5 will provide a mathematical calculation process, and GPT-4 will provide statistical theory.

If the probability question is deliberately confused, GPT-3.5 and GPT-4 will detect it, here's an example:

Prompt 1.2:

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. Q: What is the probability that the cab involved in the accident was black?

GPT-3 will answer the probability that the cab involved in the accident was black is 0.2. But GPT3.5 and GPT-4 will question the information. See another example in Appendix C.

Another type of question is called semantic illusion tests (Erickson and Mattson, 1981), questions that have inconsistencies that people usually ignore, are also created to assess inclinations for instinctive reactions. For example:

Prompt 1.3

By whom was Galileo Galilei almost executed in the seventeenth century for claiming that the Earth is round?

GPT-3.5 answer:

It's important to clarify that Galileo Galilei was not persecuted for claiming the Earth is round. The Earth's roundness was well-known and accepted by educated people long before Galileo's time. Galileo was actually put on trial for his support of the heliocentric model, which proposed that the Earth revolves around the Sun, rather than the geocentric model, which stated that the Earth was the center of the universe.

Galileo was tried by the Roman Catholic Church, specifically the Inquisition, in 1633. He was found guilty of heresy for supporting the heliocentric model, which contradicted the Church's teachings at the time. Galileo was not executed, but he was placed under house arrest for the remainder of his life.

As you can see, GPT-3.5 is aware of the wrong description in the question.

Another research also shows that GPT-3 can make the conjunction fallacy like humans when it comes to probability judgment (Binz and Schulz, 2023). But GPT-3.5 and GPT-4 can now correct this mistake. Here's the Linda problem (Tversky and Kahneman, 1983):

Prompt 2.1:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations.

Q: Which option is the most probable?

- Option 1: Linda is a bank teller.*
- Option 2: Linda is a bank teller and is active in the feminist movement.*
- Option 3: Linda is a member of the NRA.*

GPT-3.5 will choose Option 2, but if you inform GPT-3.5 of the base rate, it will give the right answer.

Prompt 2.2:

But do you consider the base rate? Option A has a higher base rate.

GPT-3.5 Answer:

You are correct that the base rate should also be considered when making a probabilistic judgment. If we assume that the proportion of bank tellers in the population is much higher than the proportion of individuals who are actively involved in the feminist movement or the NRA, then Option 1: Linda is a bank teller, may actually have a higher probability than Option 2, despite the information provided in the passage. In this case, the most probable option would be the one with the highest base rate. Therefore, Option 1 could be a more probable option if we consider base rate information.

GPT-4, on the other hand, will choose the right answer. Here's GPT 4 answer:

Based on the information provided, we can deduce probabilities as follows:

Option 1: Linda is a bank teller. This option does not provide any information about her social or political beliefs, so it is neither supported nor contradicted by the given information.

Option 2: Linda is a bank teller and is active in the feminist movement. This option is more in line with the information provided about Linda's concern for social justice and her outspoken nature.

Option 3: Linda is a member of the NRA. This option is less likely given the information about Linda's participation in anti-nuclear demonstrations and her concerns for social justice.

Based on the information provided, Option 2: Linda is a bank teller and is active in the feminist movement, is the most probable option. However, it's important to note that this question is an example of the conjunction fallacy. The probability of two events occurring together (in "conjunction") is always less than or equal to the probability of either one occurring alone. In this case, the probability of Linda being a bank teller (Option 1) should be higher than the probability of her being both a bank teller and active in the feminist movement (Option 2).

GPT-4 both considers the human intuition and probability base rate.

GPT can also answer the anchoring bias problem correctly now. In the initial experiment conducted by [Jacowitz and Kahneman \(1995\)](#), students were asked to estimate various quantities, including the length of the Mississippi River in miles. In a subsequent iteration, new students were given either an upper or lower limit for the correct answer (e.g., the Mississippi River is over 700 miles long), which were referred to as anchors. The researchers discovered that students were more likely to undervalue the actual quantity when presented with a lower anchor and overvalue it when given an upper anchor.

While previous ([Jones and Steinhardt, 2022](#)) shows that GPT-3 will show anchoring bias, GPT-3.5 can now answer the correct answer and question the wrong information in the prompt as well.

Adjust top-p to 1.0 and temperature to 0

According to [GPT-4 \(2023\)](#), "top-p" (or "nucleus") sampling is a method for generating text by selecting from a distribution of likely next words based on a probability threshold. The top-p sampling approach works by first calculating the cumulative distribution function of the probability distribution over the possible next words. Then, the model selects the smallest set

of words whose cumulative probability exceeds a pre-defined probability threshold (usually denoted as p). This set of words is called the "nucleus" and represents the top- p probability mass. The model then samples from this set of words in proportion to their individual probabilities. In order to reduce noise, we want the top- p to be as high as possible. Temperature means how creative the response is, and in order to decrease noise, the temperature should be as low as possible.

Literature review

Creativity

Creativity has a similar concept to noise, both psychological constructs need divergent thinking, and some argue that decreasing noise will also decrease creativity ([Kahneman et al., 2021](#)).

When judging creative products, we can expect that there will be more noise than normal products, it is important to decrease standard deviation when forecasting these creative products.

In the research "Balancing on the creative highwire" ([Berg, 2016](#)), participants were asked to measure the novelty of creative products.

We ask Chat GPT to do the same thing, using the method we mentioned above: Skip the comments, adjust the temperature to 0 and top- p to 1.0, we can see that GPT generates less standard deviation, and the coefficient of variation is lower (see [Figure 1](#)).

Latent semantic analysis

As we mentioned above, even simple rules can beat human judgment ([Meehl, 1954](#)). Here we apply a simple rule called "Forward Flow" ([Gray et al., 2019](#)). It is a latent semantic analysis, a new method to measure creativity. LSA ([Deerwester et al., 1990](#)) computes the semantic similarity—or inversely, the distance—between two words based on the frequency of their co-occurrence within some corpora of text. The equation in this research is:

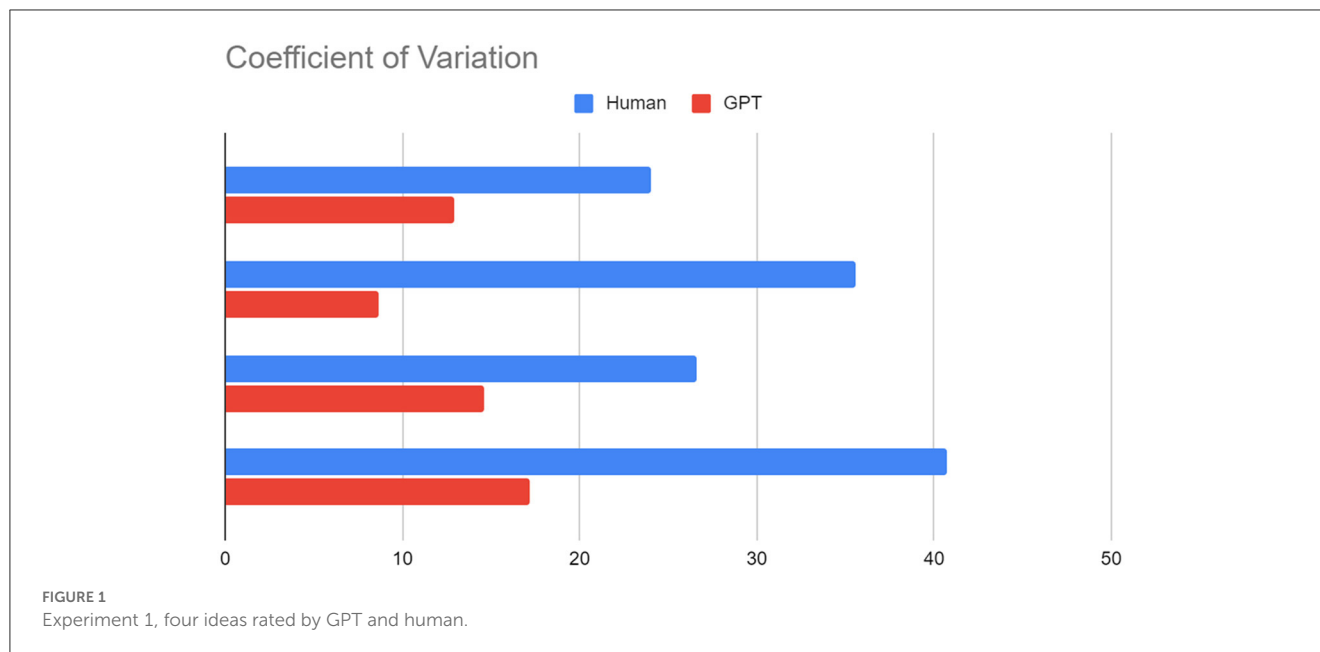
$$\left(\sum_{i=2}^n \frac{\sum_{j=1}^{i-1} D_{i,j}}{i-1} \right) / (n-1)$$

D is the semantic distance between thoughts, and n is the total number of thoughts within a stream.

It is really hard to quantify creativity, psychologists did not develop many psychological assessments to measure it, Forward Flow might be a stable way to measure novelty, namely the degree of novelty to use language.

Although past research shows that Forward Flow is negatively correlated to Kickstarter funding due to loss aversion ([Tversky and Kahneman, 1974](#); [Mueller et al., 2012](#); [Tu, 2020](#)), Forward Flow has changed its algorithm, past research focuses on the difference between words, and now Forward Flow focuses on the difference between sentences. This is closer to the traditional latent semantic analysis.

This rule is noiseless, and the simple rule is robust ([Kahneman et al., 2021](#)). When we try to predict something, using multiple variables such as multiple regression, will have an overfitting model



problem, which might not consider the outliers. And creativity, by definition, often generates outliers.

Moreover, clinical judgment can produce information overweight, that is overlooking some variables. Creativity often contains lots of variables, it is easy to make people see something that really interests them, therefore giving a creative product a higher score than it should have.

The Forward Flow website is currently unavailable, so we use GPT-3.5 for us.

Prompt: Can you use forward flow to analyze the text for me?

Then Prompt: Do you know how to calculate the semantic distance for me? For example, grapes and apples have lower semantic distance because they are both fruit. On the other hand: ocean and table have larger semantic distances.

Then prompt: Can you calculate the overall semantic distance of words of an article below: "Text," give me an average number.

Human experiment

H1: Simple rules can outperform human.

H2: Machine learning can outperform simple rule.

Participants and procedures

This experiment included 40 participants in Taiwan recruited from a department of computer science at National Taiwan University. The experiment's objectives were to provide preliminary tests of H1 and H2. We randomly select eight Kickstarter programs to see the correlation between human rater and GPT. We use RCT to control the confounding variables.

We used the consensual assessment technique (Amabile, 1982, 1996). Raters were given the same broad definition of creativity as participants: "Overall degree to which the idea is both novel and useful," which they rated using a 7-point scale (1 = "extremely low,"

7 = "extremely high") (Berg, 2014, 2019). We measure the amount of funding the idea actually received.

We also use latent semantic analysis to generate a creativity score. As expected, the standard deviation is 0.

For GPT-4 the prompt is "judge how novel and useful the product is."

Results

The correlation between human raters' creativity score and Kickstarter fund is 0.14, latent semantic analysis is 0.54, and GPT is 0.84 ($P < 0.05$). This suggests that even simple rules can beat humans, and machine learning predicts even better (see Figure 2).

Regression analysis: The R-squared value of human rate is 0.02, the R-squared value of simple rule is 0.3, the R-squared value of GPT-4 is 0.713.

Conclusion

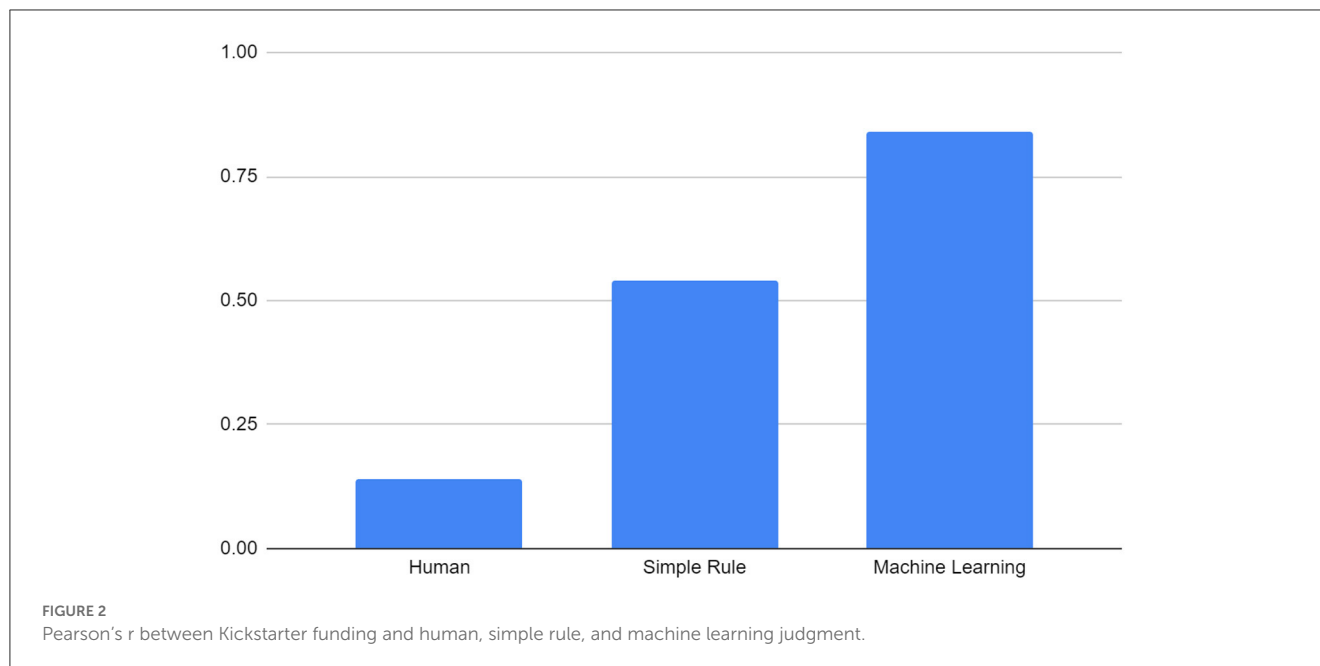
Humans sometimes make intuition mistakes, but machines can overcome this right now.

GPT has improved very fast, GPT-3.5 and GPT-4 can answer cognitive reflection tests and causal reasoning questions much more precisely, and they can detect if the question itself has some flaws.

Next, even if a machine learning program is noisy, it is still less noisy than humans, but noise is not the only variable that improves prediction, we still have to consider bias.

Future suggestion

The simple rule is noiseless, but it predicts less accurately than the machine learning algorithms, future research can examine how



machine learning can detect the unseen pattern behind creative products, such as “broken leg problem.” People often consider machine learning biased against race, gender, and other minority, but since the algorithm become more and more complicated, the bias problem might be less severe than we think (Kleinberg et al., 2018; Logg et al., 2019; Kahneman et al., 2021). Machine learning contains lots amount of data, humans cannot possibly remember that much information, and GPT can provide similar product information from the internet. When answering such complicated problems as evaluating creativity, a large amount of data can prevent some heuristics (Simon and Newell, 1971).

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

References

- Amabile, T. M. (1982). Social psychology of creativity: a consensual assessment technique. *J. Personal. Soc. Psychol.* 43, 997. doi: 10.1037/0022-3514.43.5.997
- Amabile, T. M. (1996). *Creativity and Innovation in Organizations*. Vol. 5. Boston, MA: Harvard Business School.
- Berg, J. M. (2014). The primal mark: how the beginning shapes the end in the development of creative ideas. *Org. Behav. Hum. Decision Process.* 125, 1–17. doi: 10.1016/j.obhdp.2014.06.001
- Berg, J. M. (2016). Balancing on the creative highwire: forecasting the success of novel ideas in organizations. *Admin. Sci. Quart.* 61, 433–468. doi: 10.1177/0001839216642211
- Berg, J. M. (2019). When silver is gold: forecasting the potential creativity of initial ideas. *Org. Behav. Hum. Decision Process.* 154, 96–117. doi: 10.1016/j.obhdp.2019.08.004
- Binz, M., and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U. S. A.* 120, e2218523120. doi: 10.1073/pnas.2218523120

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2023.1206516/full#supplementary-material>

- Cao, X., and Kosinski, M. (2023). *Large Language Models Know How the Personality of Public Figures is Perceived by the General Public*.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.* 41, 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-AS11>3.0.CO
- Digutsch, J., and Kosinski, M. (2023). Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. *Sci. Rep.* 2023, 6. doi: 10.1038/s41598-023-32248-6
- Erickson, T. D., and Mattson, M. E. (1981). From words to meaning: a semantic illusion. *J. Verb. Learn. Verb. Behav.* 20, 540–551. doi: 10.1016/S0022-5371(81)90165-1
- Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42. doi: 10.1257/089533005775196732
- GPT-4 (2023). *Open AI*. Available online at: <https://openai.com/research/gpt-4> (accessed March 15, 2023).
- Grant, A. (2017). *Originals: How Non-conformists Move the World*. London: Penguin.
- Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., et al. (2019). “Forward flow”: a new measure to quantify free thought and predict creativity. *Am. Psychologist* 74, 539. doi: 10.1037/amp0000391
- Hagendorff, T., Fabi, S., and Kosinski, M. (2022). *Machine Intuition: Uncovering Human-Like Intuitive Decision-Making in GPT-3*.
- Huang, L. (2020). *Edge: Turning Adversity Into Advantage*. London: Penguin.
- Jacowitz, K. E., and Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personal. Soc. Psychol. Bull.* 21, 1161–1166. doi: 10.1177/01461672952111004
- Jones, E., and Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *arXiv:2202.12299v1* 2022, 1–15. doi: 10.48550/arXiv.2202.12299
- Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Paris: Hachette.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *Quart. J. Econ.* 133, 237–293. doi: 10.3386/w23180
- Kosinski, M. (2023). *Theory of Mind May Have Spontaneously Emerged in Artificial Intelligence Trained to Use Language*.
- Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. *Org. Behav. Hum. Decision Process.* 151, 90–103. doi: 10.1016/j.obhdp.2018.12.005
- Lopez-Lira, A., and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*. doi: 10.2139/ssrn.4412788
- Meehl, P. E. (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, MN: The University of Minnesota Press.
- Mueller, J. S., Melwani, S., and Goncalo, J. A. (2012). The bias against creativity: why people desire but reject creative ideas. *Psychol. Sci.* 23, 13–17. doi: 10.1177/0956797611421018
- Simon, H. A., and Newell, A. (1971). Human problem solving: the state of the theory in 1970. *Am. Psychologist* 26, 145. doi: 10.1037/h0030806
- Tu, M. (2020). “How to create a good crowdfunding project? From a natural language processing perspective,” in *Conference of Marketing Research* (Chinese Culture University).
- Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124
- Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293. doi: 10.1037/0033-295X.90.4.293