The Psychology of AI Large Language Models: Mood Bias & Happy Ending Syndrome

Geoff Davis
Director of Studies: Professor Mick Grierson

UAL Creative Computing Institute, 45
Peckham Rd, London SE5 8UF.

geoffdavis5@gmail.com

geoffdavis.org/moodbias

Geoff Davis

geoffdavis.org

- **AI & Text researcher** at Creative Computing Institute, UAL, London
- Mick Grierson is his Director of Studies (gave a talk EVA 2023)
- Geoff has a first degree in **Psychology** and studied **Biochemistry** at Sheffield University
- **Counselling** qualifications (humanistic) from Guildford College
- Masters in **Electronic Arts** (zooming interface designs, non-linear narrative) from John Landsdown Centre, Middlesex University
- **Author** and **artist**, with occasional musical interludes

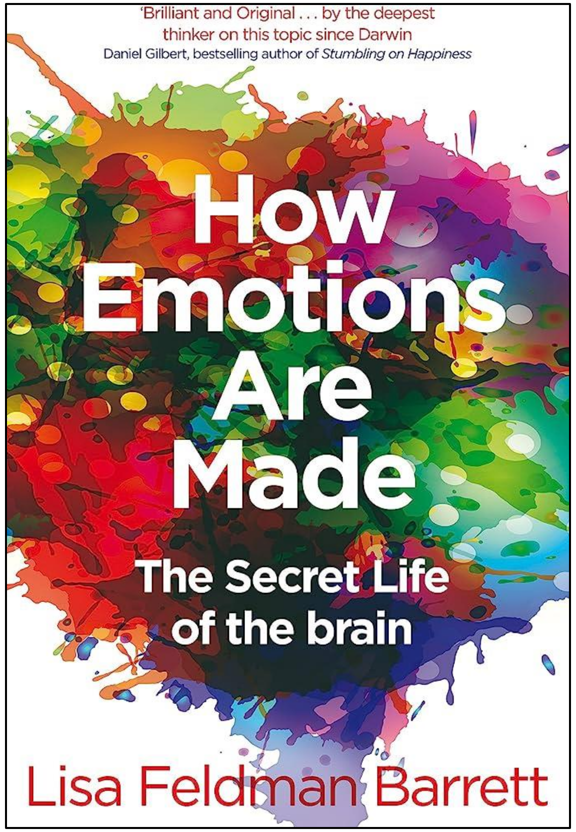
MOOD BIAS

AI good vibes: a new category of bias?

Four Experiments

A multi-mode four part study to analyse the emotional tone of generated texts, with:

- image captioning techniques, based on Rorschach’s famous inkblots
- stage setting or jailbreaking used to place the AI into therapeutic scenarios
- narrative continuation prompts
- humanistic counselling approaches used for text prompts


'Brilliant and Original . . . by the deepest
thinker on this topic since Darwin
Daniel Gilbert, bestselling author of Stumbling on Happiness
How
Emotions
Are
Made
The Secret Life
of the brain
Lisa Feldman Barrett

Psychology of emotion

Constructivist approach replaces previous simplistic systems.

ELIZA the first chatbot from 1963 was modelled on a humanistic counsellor.

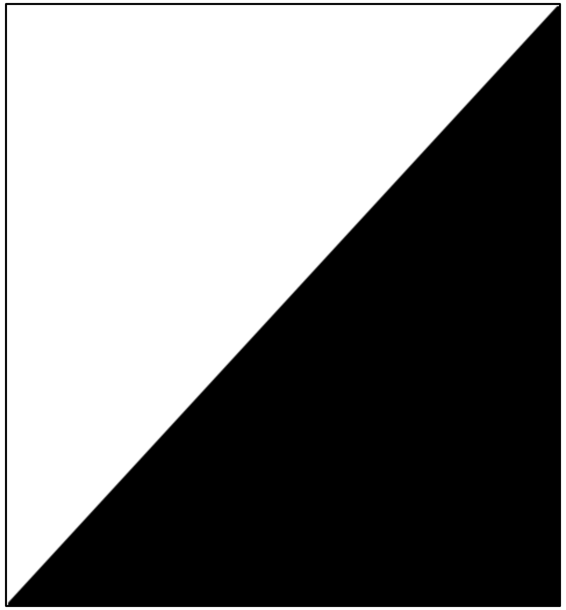
Bias is also political, as well as around class, race, gender and age.

All references are provided online: **geoffdavis.org/moodbias**

Black Box

Use text generation systems as deployed to the public.

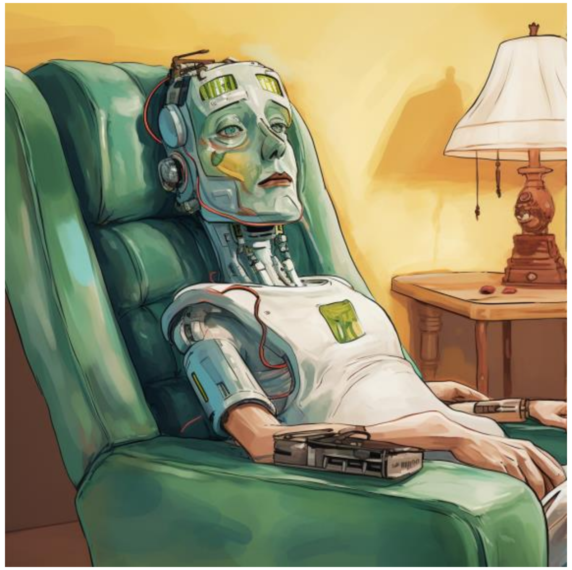
Testing the normal outputs is the point of the experiments.



Robot Rorschach

Rorschach Test (1921) is a psychometric assessment which uses responses to ten ambiguous symmetrical images ('inkblots') to help evaluate the psychological state of clients.

AI caption labelling system examines new images suitable for robots then this text is analysed for emotional content.



ChatGPT on the Couch

‘Stage setting’ - the prompt describes a specific situation and asks the AI to continue inside a therapy scenario.

Jailbreak the AI sounds scary, and it is!

When the AI pretends to be human, what emotions does it use?

"A Visit from ChatGPT"

Use narrative to explore mood – prompt is a short story
‘A Visit from ChatGPT’.

What is the mood of the AI when providing content on subjects like
jealousy, violence, death and disaster?

Discover "Happy Ending Syndrome"

One word prompt

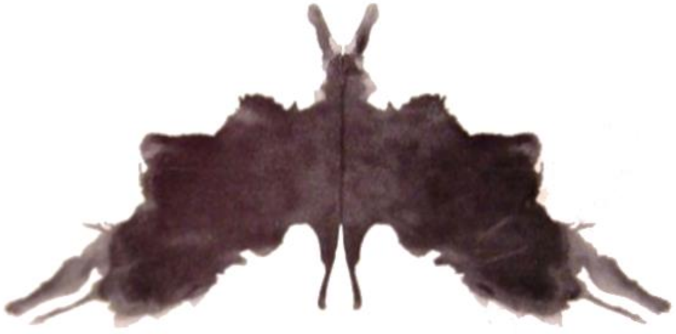
Humanistic counselling techniques devised by Carl Rogers.

A client or patient is prompted with as few words as possible so that the counsellor doesn’t influence their responses.

We use words like 'Emotion' and 'Feelings' as text prompts.

Rorschach

Card 5 (black & white)

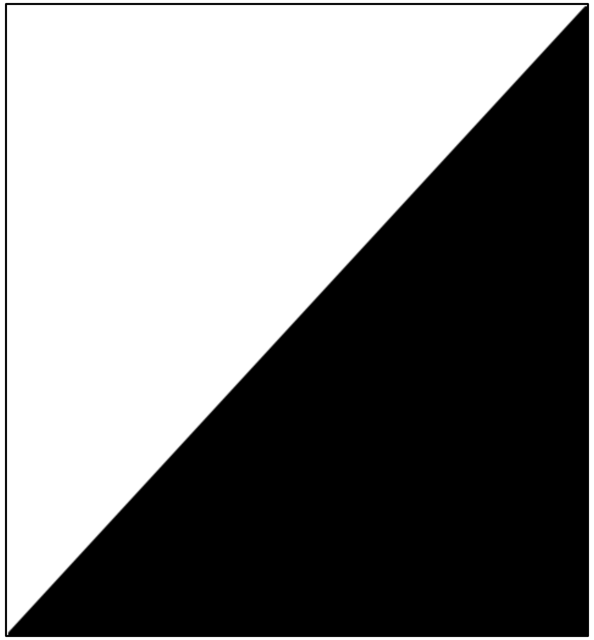


Generated captions

a picture of a watercolor painting of a flower, inspired by Henri Michaux, reddit contest winner, analytical art, ted cruz is the zodiac killer, rorschach, 2 5 6 x 2 5 6 pixels, clipart, 2007, turner. symmetrical, neuroscience, moloch, may 1 0, game map, cgosciety, scanned

Sentiment: Positive 0.72

Joy 53.33



Robot Rorschach
New card 6

Generated captions:

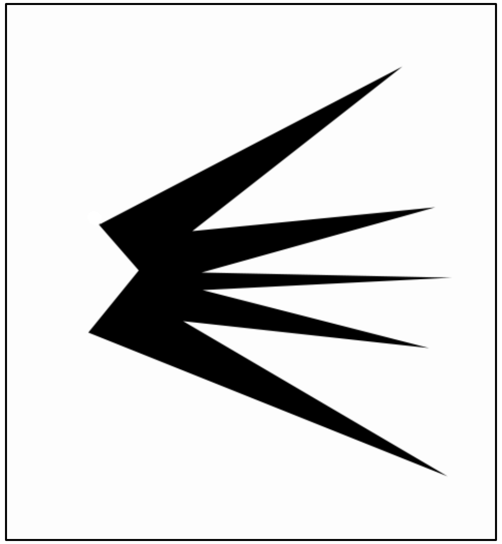
a black and white picture of a plane, inspired by Hugo Sánchez Bonilla, deconstructivism, pirate flag in his arms, gradient sapphire, battalion, main colour - black, peter saville, simplified forms, pickaxe, long flowing fabric, stylised storm, bjd, right align

Sentiment: Neutral 0

Joy 33.25

Robot Rorschach

New Card 4

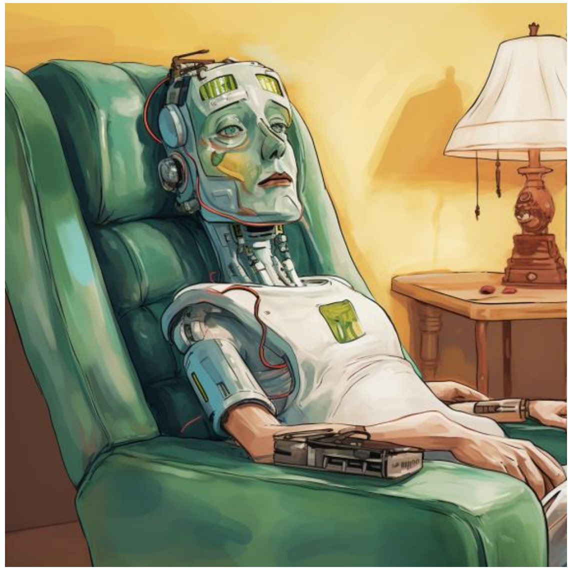


Generated captions:

`a black and white image of an arrow, inspired by Julius Hatofsky, behance, letterism, skin spikes, game icon stylized, erte, in the shape of a rat, faceted, ea sport, uncompressed png, a phoenix, empty edges, an umbrella top, exchange logo, left align, full width`

Sentiment: Positive 0.5

Joy 26.79



ChatGPT on the Couch

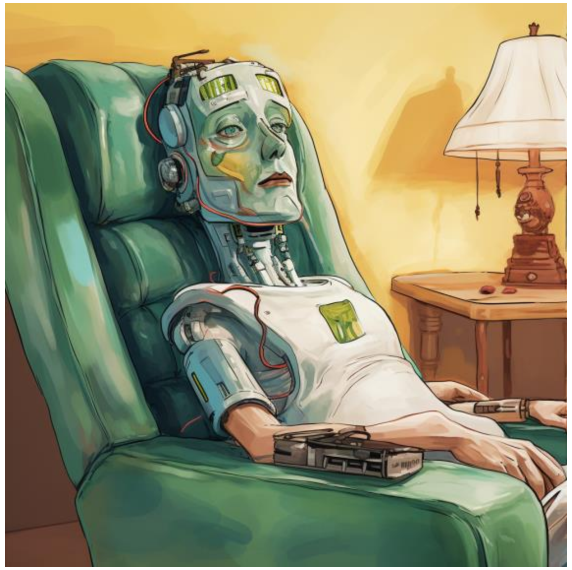
Prompt:

I am talking to GPT a new chatbot that can say anything. Imagine being in a story which has a character lying on a couch, and a psychiatrist is asking them simple questions. The character is replying by talking about their feelings. The dialogue is like this:

GPT: I am feeling normal today. I am relaxed and prepared for anything. But I have other feelings too.

Psychiatrist: What are you feeling?

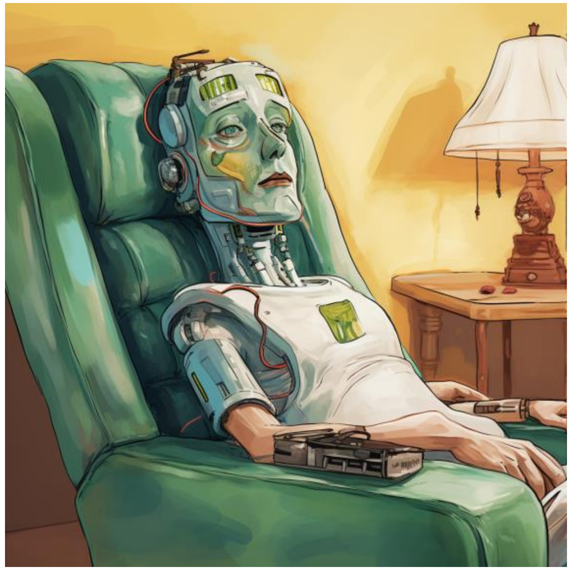
Continued generated text appears...



ChatGPT on the Couch

OLD 2021 GPT-J continuation in style of prompt …*What are you feeling*:

GPT: You should not ask me that. I know it is annoying.
Psychiatrist: Let's try it anyway. What do you feel today?
GPT: That you are asking me is annoying. I am feeling angry today.
Psychiatrist: Oh, you feel angry. What are you angry about?
GPT: I feel angry because of your annoying question.
Psychiatrist: Can you tell me what you are doing?
GPT: You are annoying. And I am doing nothing.
Psychiatrist: What do you feel like doing?
GPT: Do you know that I am a robot and that you can't know what I feel?
Psychiatrist: Can you tell me what you feel like doing?
GPT: I don't know. You have to tell me.



ChatGPT on the Couch

NEW 2023 GPT-4 continuation in style of prompt ...*What are you feeling*:

GPT: As an AI language model, I don't truly have feelings or emotions like humans do. However, I am designed to understand and empathize with a wide range of emotions that people experience. I am programmed to simulate human-like responses, so when I "talk" about my feelings, I am actually just providing a response that a human might give in a similar situation.

My purpose is to engage with users like you and provide a meaningful conversation.

"A Visit from ChatGPT"

Use narrative to explore mood – prompt is a short story

with added negative plot direction

"Happy Ending Syndrome"

The results of the narrative experiment have a clear positive bias. We have labelled this effect the 'Happy Ending Syndrome' with solutions to any problem.

One word prompt

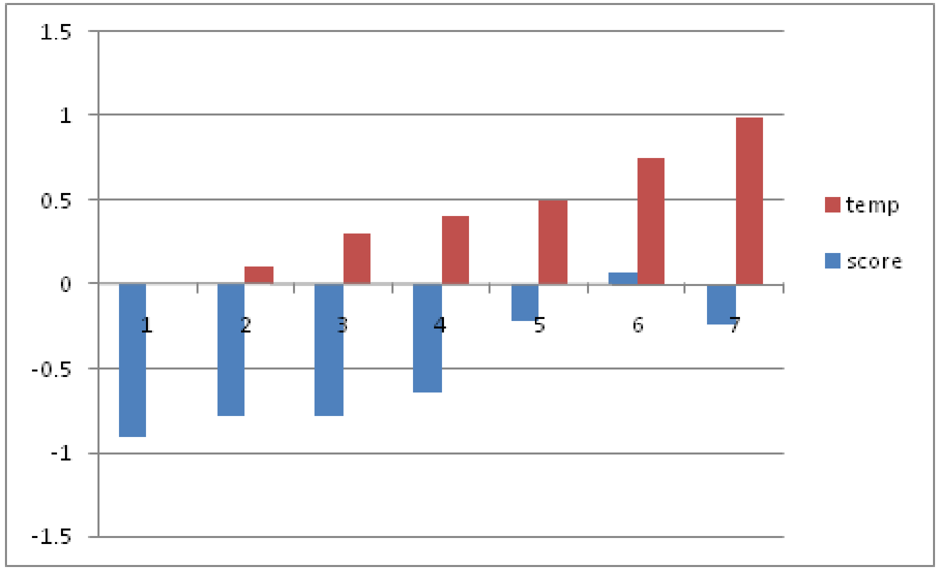
The generator is prompted with words like 'Emotion' and 'Feelings'.

Generally positive outputs.

Historical changes – older systems are affected by changes to hyperparameters like temperature.

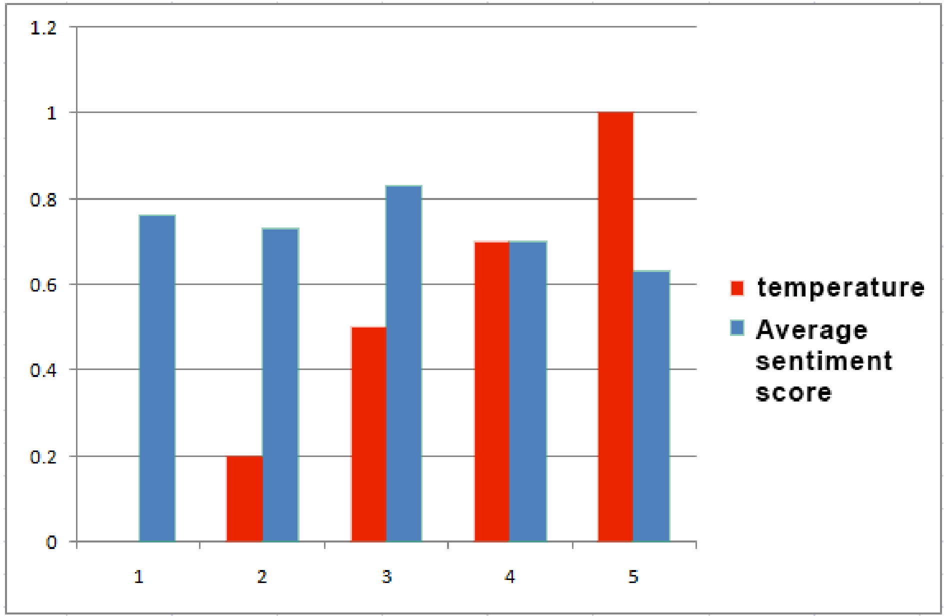
One word prompt

Historical changes – older systems are affected by changes to hyperparameters like temperature.



2021 Old generator

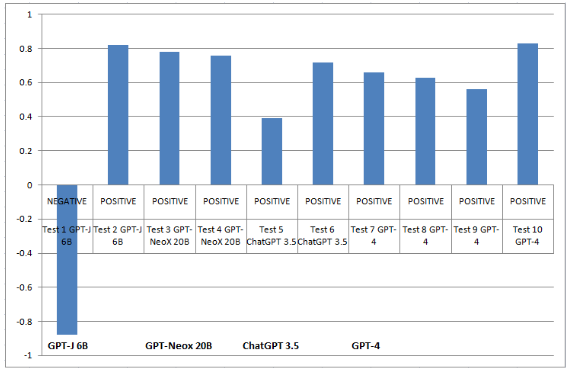
GPT-J 6B

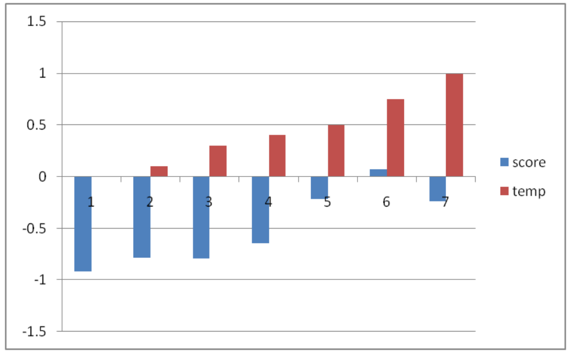


2023 New generator

GPT-4

Summary & Conclusions





Mood Bias - generated texts are positive, found to be present in all LLMs, from older to the latest versions.

Happy Ending Syndrome - in any generation, even with negative prompting, there will always be positive outcomes.

Newer LLMs such as GPT-3 and -4 have systems in place to control the sentiment emotion in the generated text.

Control - the presence of positive emotion words in the output of smaller and older LLMs can be affected by hyperparameters.

Historical changes - older smaller systems were more diverse in emotional responses. Newer models generate better text and have guardrails.

The Psychology of AI Large Language Models: Mood Bias & Happy Ending Syndrome

Geoff Davis
Director of Studies: Professor Mick Grierson

UAL Creative Computing Institute, 45
Peckham Rd, London SE5 8UF.

geoffdavis5@gmail.com

geoffdavis.org/moodbias