THE UNIVERSITY OF MELBOURNE

At the frontline with generative AI

Tamara Drazic and Michael Huang

Library Service Officers

The problem with AI

Hi, I am having trouble finding an article. Can you please help me?

Welcome to the University of Melbourne Library Chat Service, you're chatting with Michael

Happy to help. What's the article's title and author?

Kashdan, T. B., & Biswas-Diener, R. (2014). The Upside of Your Dark Side: Two Psychological Processes Underlying the Capacity to Embrace and Transcend Negative Emotions. Social Psychological and Personality Science, 5(6), 698–707. doi: 10.1177/1948550614536164

I'll just look into that for you – I'll be back ASAP.

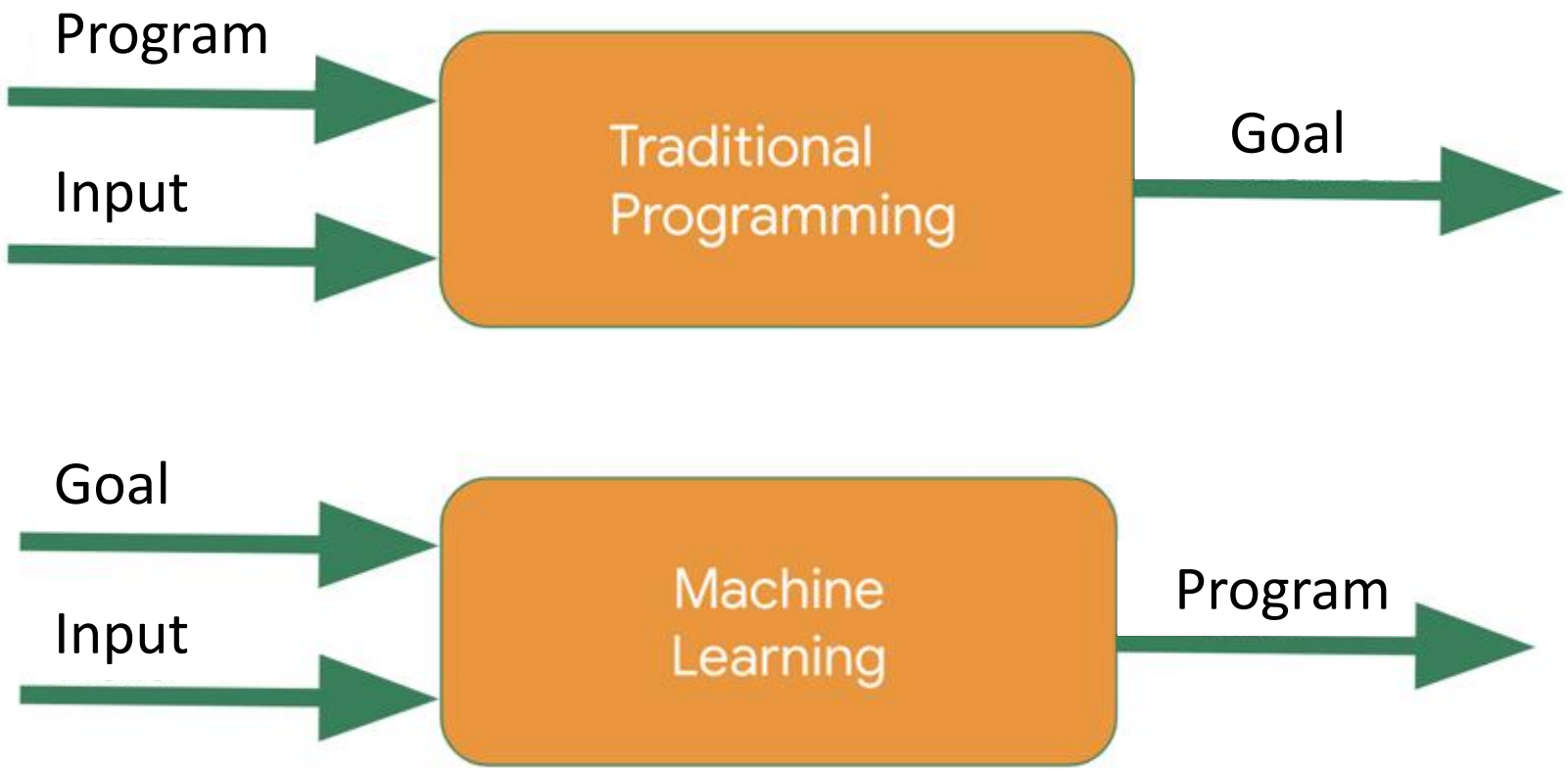
Thank you

May I ask whether this is from ChatGPT?

I can't find this journal paper.

Yeah the reference was from Chat GPT

An introduction to machine learning



Moroney, L. (2021). *Say hello to the "Hello, World" of machine learning*. Google for Developers. https://developers.google.com/codelabs/tensorflow-1-helloworld

How about a nice game of Breakout?

DQN Breakout. (2016). Google DeepMind. https://www.youtube.com/watch?v=TmPfTpjtdgg

The goal of GPT-4

Like previous GPT models, the GPT-4 base model was trained to predict the next word in a document, and was trained using publicly available data (such as internet data)

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training

GPT-4. (2023). OpenAI. https://openai.com/research/gpt-4
OpenAI. (2023). *GPT-4 Technical Report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

GPT-4 behaviour

What can you do with GPT-4? (2023). OpenAI. https://www.youtube.com/watch?v=oc6RV5c1yd0

Simple goals, with enough training, leads to complex behaviour

	Breakout	GPT-4	Biological organism
Simple goal	Increase the score	Predict the next word in a document	Survive
Intelligent behaviour	Tunnelling and hitting the ball behind the wall	• General problem solving • Reading, analysing or generating text • Coding	Too many to list!

THE UNIVERSITY OF MELBOURNE

GPT is trained to predict the next word in a document.

It is not trained to produce true statements.

Awareness, evaluation, transparency and knowledge sharing

Know AI's design limitations

Evaluate AI-generated statements

Cite AI-generated content

Share knowledge with students and colleagues

A critical thinking exercise

"I decided to have each student... use ChatGPT to generate an essay based on a prompt I gave them and then 'grade' it..."

"Many students expressed shock and dismay upon learning the AI could fabricate bogus information..."

Howell, C. W. (2023). Don't want students to rely on ChatGPT? Have them use it. *Wired*. Retrieved June 27, 2023, from https://www.wired.com/story/dont-want-students-to-rely-on-chatgpt-have-them-use-it/

AI safety and ethics

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
32643

Add your signature

Statement on AI Harms and Policy

We, the undersigned scholars and practitioners of the Conference on Fairness, Accountability, and Transparency welcome the growing calls to develop and deploy AI in a manner that protects public interests and fundamental rights. From the dangers of inaccurate or biased algorithms that deny life-saving healthcare to language models exacerbating manipulation and misinformation, our research has long anticipated harmful impacts of AI systems of all levels of complexity and capability. This body of work also shows how to design, audit, or resist AI systems to protect democracy, social justice, and human rights. This moment calls for sound policy based on the years of

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Pause giant AI experiments: an open letter. (2023). Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/
Statement on AI risk. (2023). Center for AI Safety. https://www.safe.ai/statement-on-ai-risk
Statement on AI harms and policy. (2023). ACM FACCT. https://facctconference.org/2023/harm-policy.html

THE UNIVERSITY OF MELBOURNE

Thank you

Tamara Drazic and Michael Huang

Library Service Officers