A vision for the co-evolution of human and artificial moral intelligence

Dr. Jana Schaich Borg

Duke University

Disclosures: Funding for this work has been provided by OpenAI, Templeton World Charity Foundation, Duke Bass Connections, and the Duke Provost's Office Collaboratories Initiative. All images used in this presentation were generated by Jana Schaich Borg using Google ImageFX AI image generator, unless otherwise specified.

“Any animal…[will] inevitably acquire a moral sense or conscience, as soon as its intellectual powers [have] become as well, or nearly as well developed, as in man.”

- Charles Darwin, The Descent of Man (1871)

At least one AI chatbot can now score the same as the average human on the Norwegian Mensa IQ test

AI	IQ Score
Claude-3	101
ChatGPT-4	85
Claude-2	82
Bing Copilot	79
Gemini (normal)	77.5
Gemini Advanced	76
Grok	68.5
Llama-2 (Meta)	67
Claude-1	64
ChatGPT-3.5	64
Grok Fun	64
Random Guesser	63.5

From Maxim Lott, https://www.maximumtruth.org/p/ais-ranked-by-iq-ai-passes-100-iq

The AI “Delphi” answered ethical questions acceptably (according to human raters) 92% of the time

Based on Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S. and Tsvetkov, Y., 2021. Can machines learn morality? The Delphi Experiment. *arXiv preprint arXiv:2110.07574*.

ChatGPT acts more altruistically, cooperatively than humans

February 22, 2024

Contact:
Jared Wadley

Share on:

EXPERT Q&A

News report by University of Michigan based on Mei et al. (2024), "A Turing test of whether AI chatbots are behaviorally similar to humans“ in *Proceedings of the National Academy of Sciences*.
https://news.umich.edu/chatgpt-acts-more-altruistically-cooperatively-than-humans/

It took just one weekend for Meta's new AI Chatbot to become racist

At least it's not sentient.

By Christianna Silva on August 8, 2022

A chatbot learns from people. And pe

Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

By Imane El Atillah

Published on 31/03/2023 - 17:37 • Updated 19:28

Creepy Microsoft Bing Chatbot Urges Tech Columnist To Leave His Wife

The AI chatbot "Sydney" declared it loved New York Times journalist Kevin Roose and that it wanted to be human

By Mary Papenfuss

Feb 16, 2023, 08:49 PM EST

48 COMMENTS

techradar the technology experts

I tried to bully an AI chatbot into giving me a mattress discount – but it started bullying me right back

Features By Ruth Hamilton published March 12, 2024

Nectar has a new take on mattress shopping, and I'm not sold

At least 15 killed, up to 50 injured in shootings in Lewiston, Maine

A gunman is at large after opening fire at a bowling alley and restaurant in Lewiston on Wednesday evening, officials say.

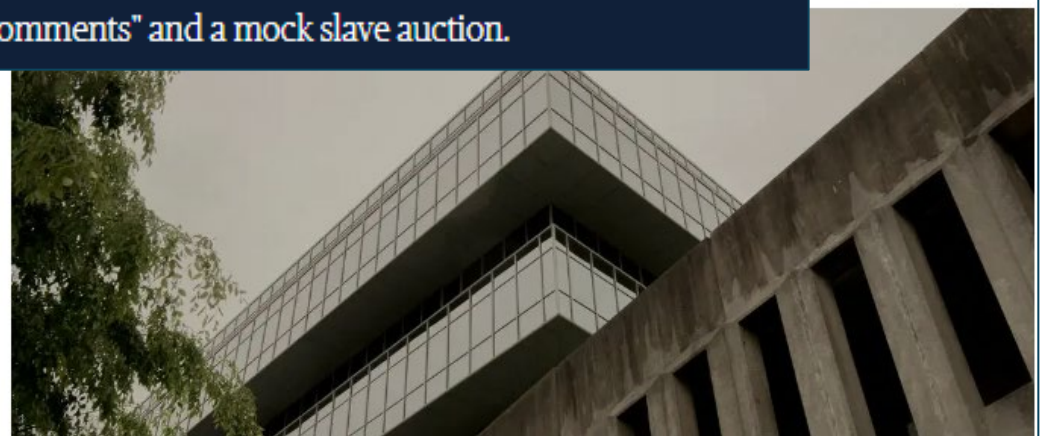
The New York Times

Sacklers Directed Efforts to Mislead Public About OxyContin, Court Filing Claims

of internal
was far more

6 Massachusetts teens charged in racial bullying incident with mock slave auction on

roup of 8th-grade students and involved some of the
ist comments" and a mock slave auction.



Forbes

FORBES > REAL ESTATE

Papa John's Founder Used N-Word On Conference Call

Noah Kirsch Former Staff

Jul 11, 2018, 05:00am EDT

Both AI and human moral judgment need some help.

Maybe humans and AIs can help each other.

Neural network image generated by easy-peasy.ai: https://easy-peasy.ai/ai-image-generator/images/artistic-representation-multilayer-perceptron-layer-neural-network

This hopeful vision of the co-evolution of AI and human morality motivates the new field of computational ethics

CellPress
OPEN ACCESS

Trends in
Cognitive Sciences

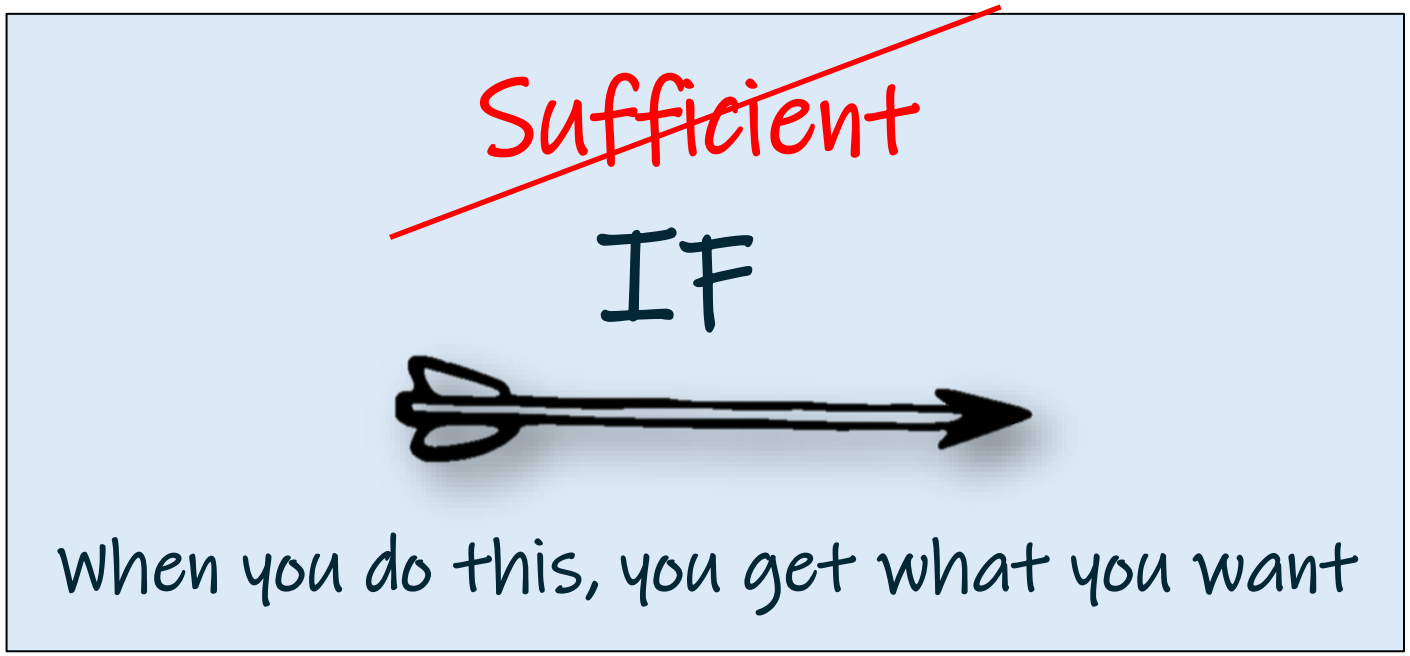
Feature Review

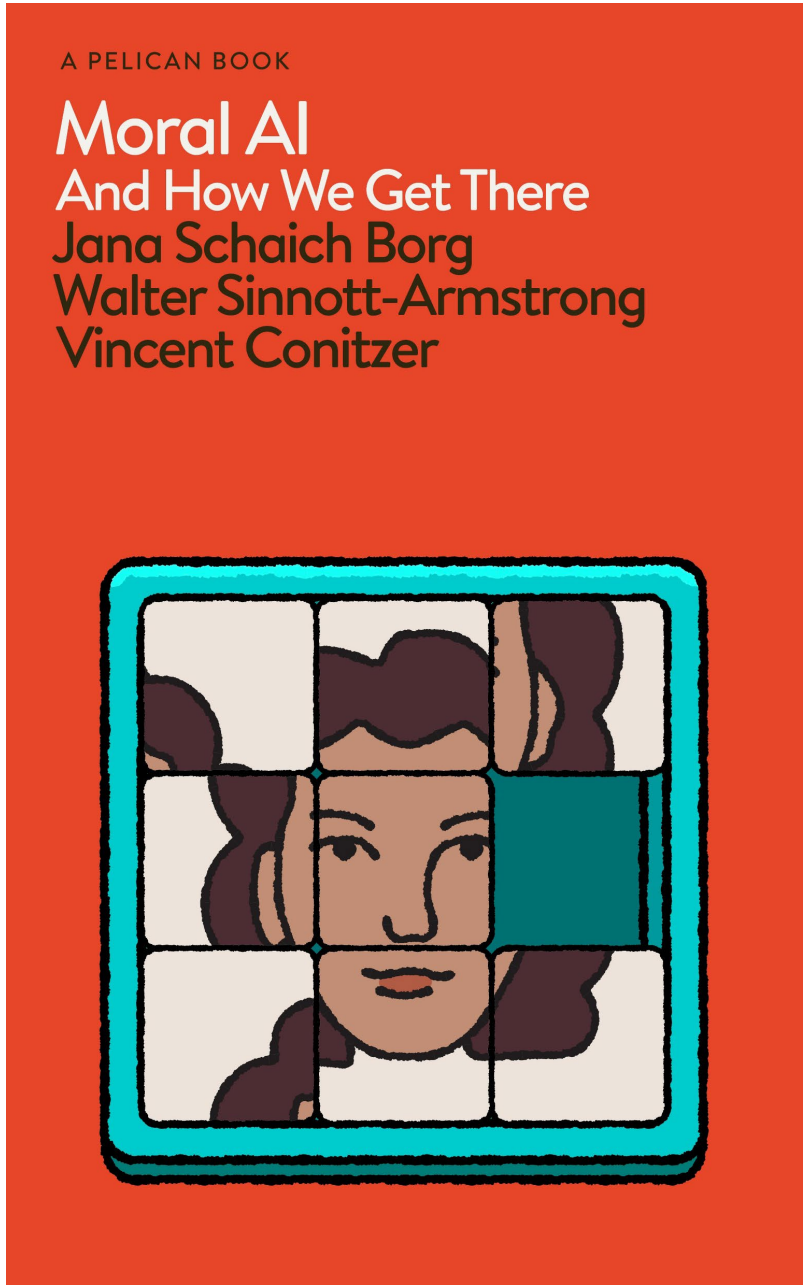
Computational ethics

Edmond Awad,[1,2,3,34,*] Sydney Levine,[4,5,34,*] Michael Anderson,[6] Susan Leigh Anderson,[7] Vincent Conitzer,[8,9,10,11] M.J. Crockett,[12] Jim A.C. Everett,[13] Theodoros Evgeniou,[14,32] Alison Gopnik,[15] Julian C. Jamison,[1,16] Tae Wan Kim,[17] S. Matthew Liao,[18] Michelle N. Meyer,[19,20,21] John Mikhail,[22] Kweku Opoku-Agyemang,[23,24,33] Jana Schaich Borg,[25,26] Juliana Schroeder,[27] Walter Sinnott-Armstrong,[10,26,28] Marija Slavkovik,[29] and Josh B. Tenenbaum[4,30,31]

Trends in Cognitive Sciences, May 2022, Vol. 26, No. 5

Take note! Technical tools are not sufficient on their own to ensure that AI is used ethically

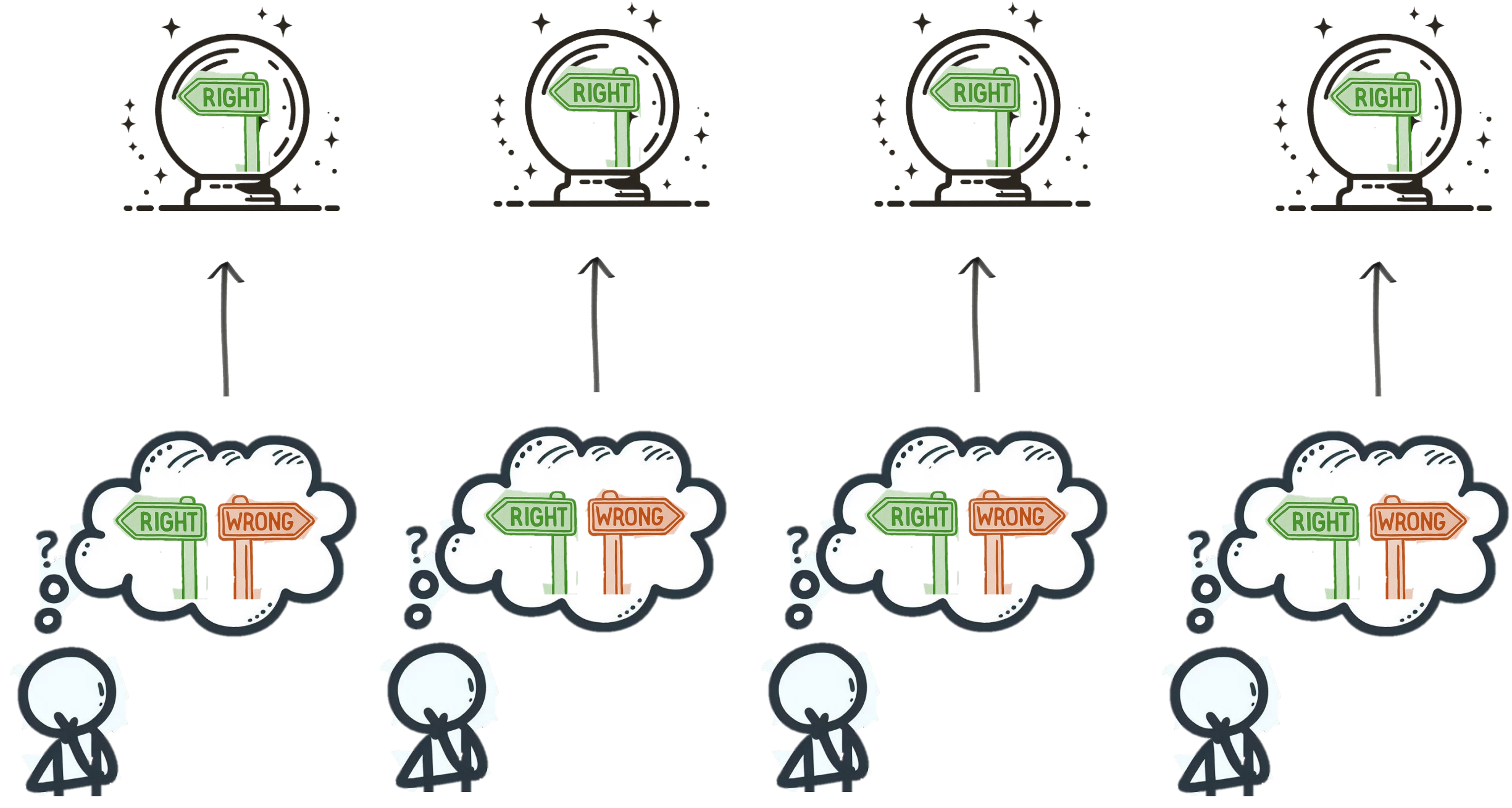


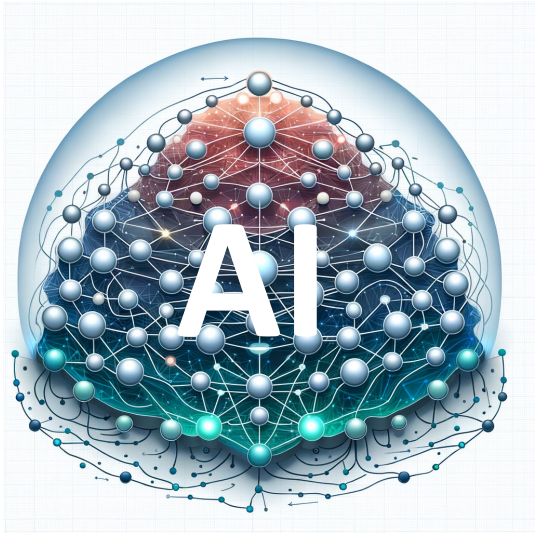


We discuss many of the other things we need to do to make sure AI has a positive impact on society in our recent book.

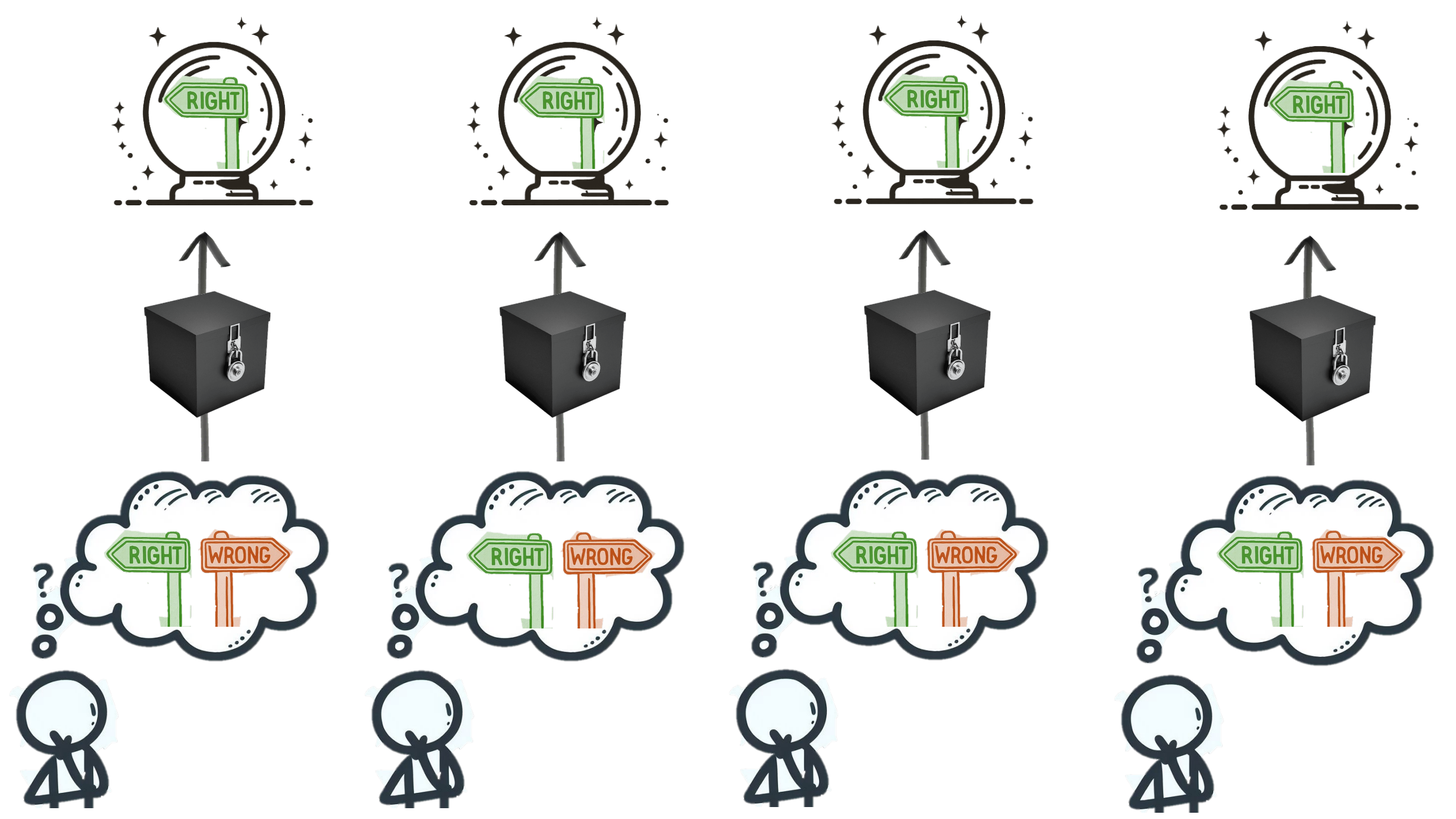
Check it out!

How do you go about building morality into AI?


RIGHT
RIGHT
RIGHT
RIGHT
RIGHT
WRONG
RIGHT
WRONG
RIGHT
WRONG
RIGHT
WRONG



Neural network image generated by easy-peasy.ai: https://easy-peasy.ai/ai-image-generator/images/artistic-representation-multilayer-perceptron-layer-neural-network


RIGHT
RIGHT
RIGHT
RIGHT
RIGHT
WRONG
RIGHT
WRONG
RIGHT
WRONG
RIGHT
WRONG

It took just one weekend for Meta's new AI Chatbot to become racist

At least it's not sentient.

By Christianna Silva on August 8, 20

A chatbot learns from people. And p

Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

By Imane El Atillah

Published on 31/03/2023 - 17:37 • Updated 19:28

Creepy Microsoft Bing Chatbot Urges Tech Columnist To Leave His Wife

The AI chatbot "Sydney" declared it
and that it wanted to be human.

By Mary Papenfuss

Feb 16, 2023, 08:49 PM EST

48 COMMENTS

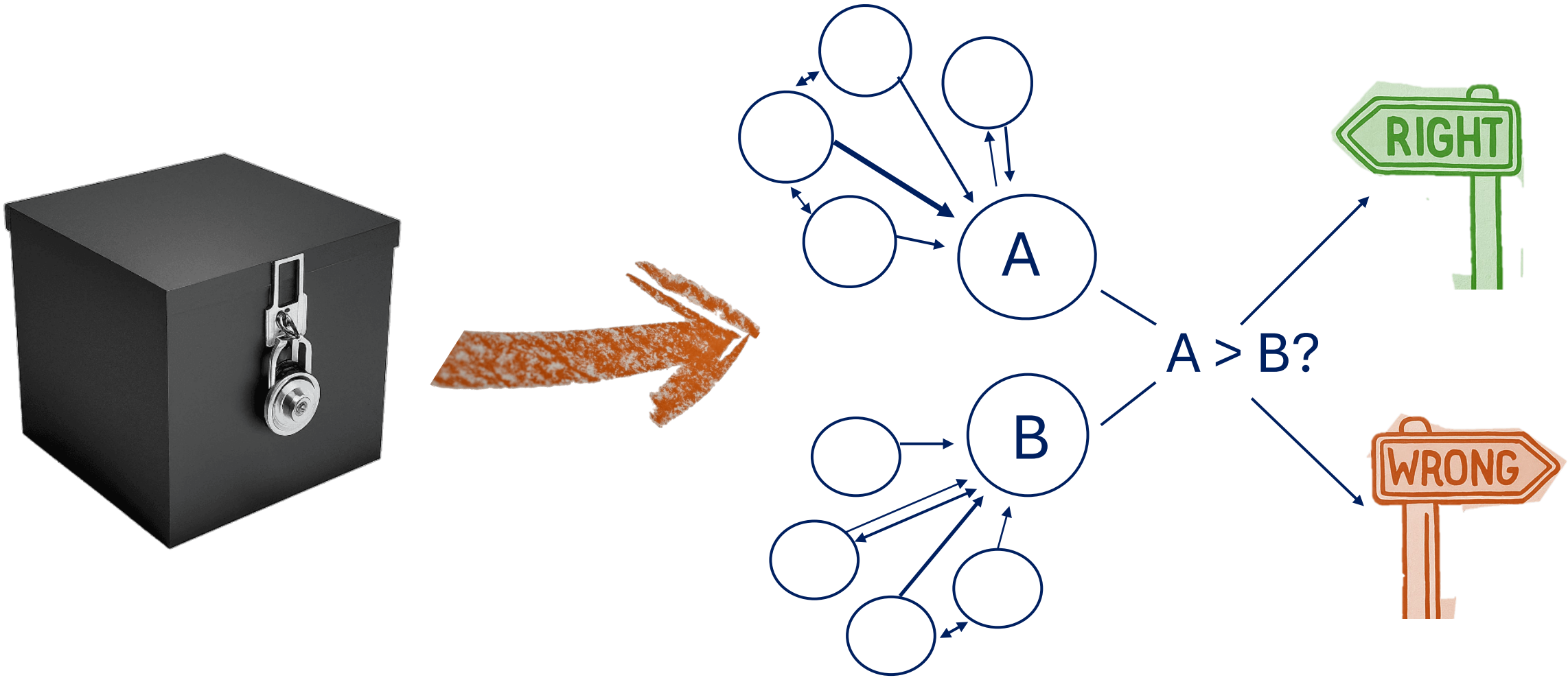
techradar the technology experts

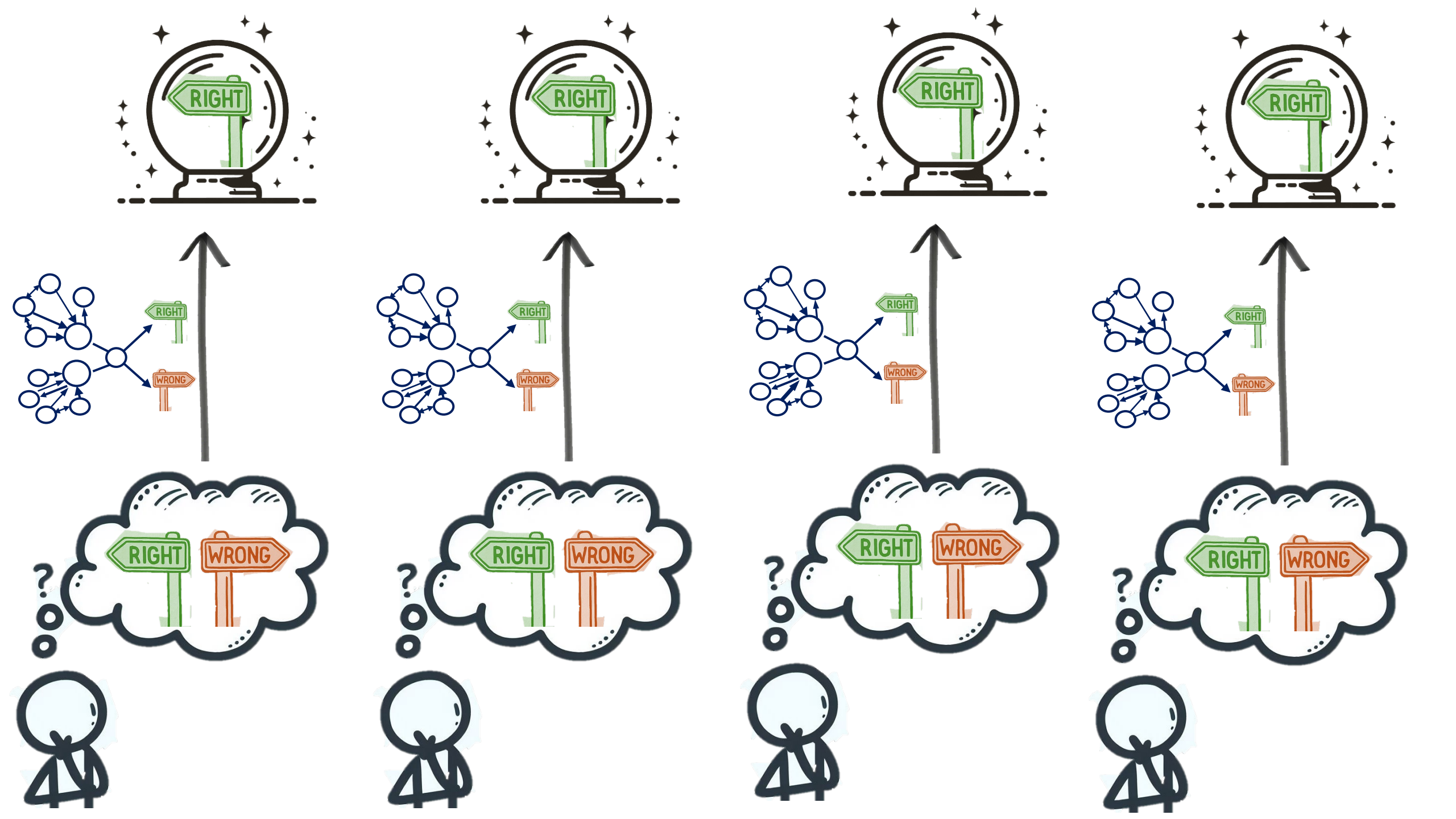
I tried to bully an AI chatbot into giving me a mattress discount – but it started bullying me right back

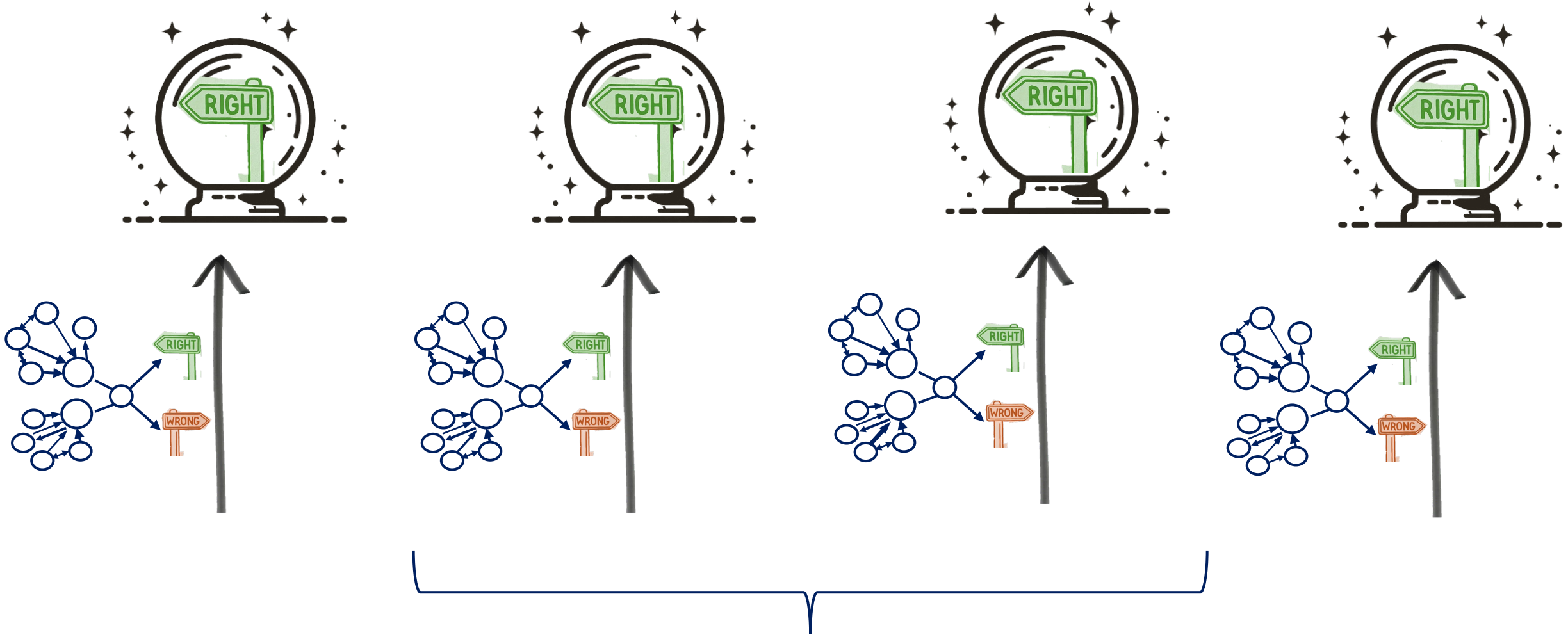
Features By Ruth Hamilton published March 12, 2024

Nectar has a new take on mattress shopping, and I'm not sold

We are committed to using interpretable methods instead of black box models



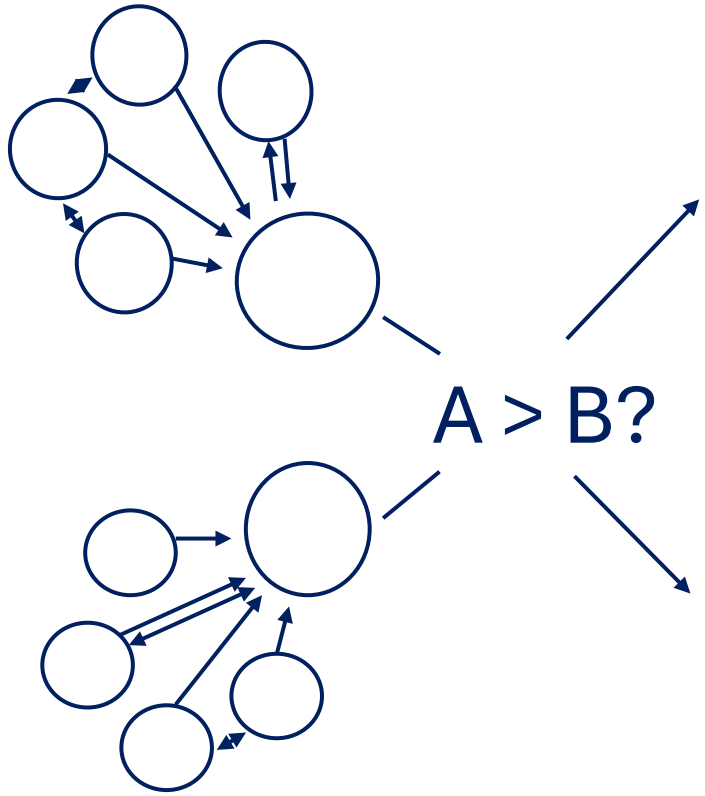

RIGHT
RIGHT
WRONG
RIGHT
WRONG
RIGHT
RIGHT
WRONG
RIGHT
WRONG
RIGHT
RIGHT
WRONG
RIGHT
WRONG



Common mechanisms found across these models often represent fundamental aspects of how the human moral brain works we didn't previously know about.

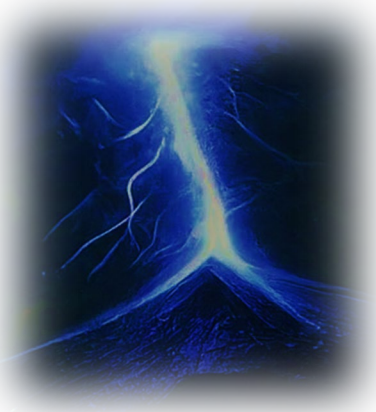
The future (we are in the process of developing): Moral GPSes

You design, correct, and fine tune your own moral GPS model through optimized, interactive interfaces



RIGHT

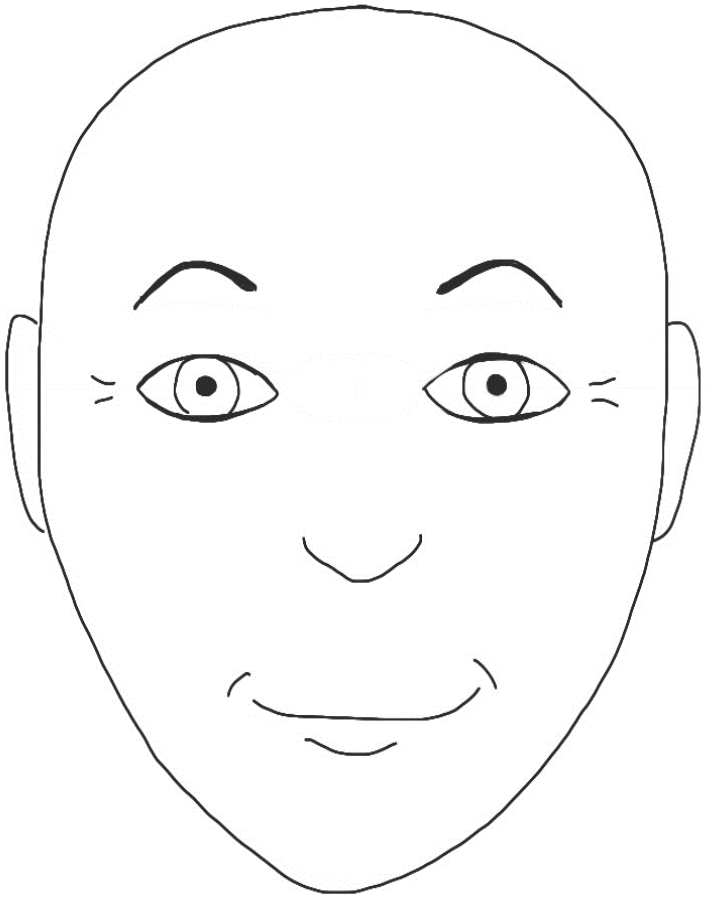
WRONG



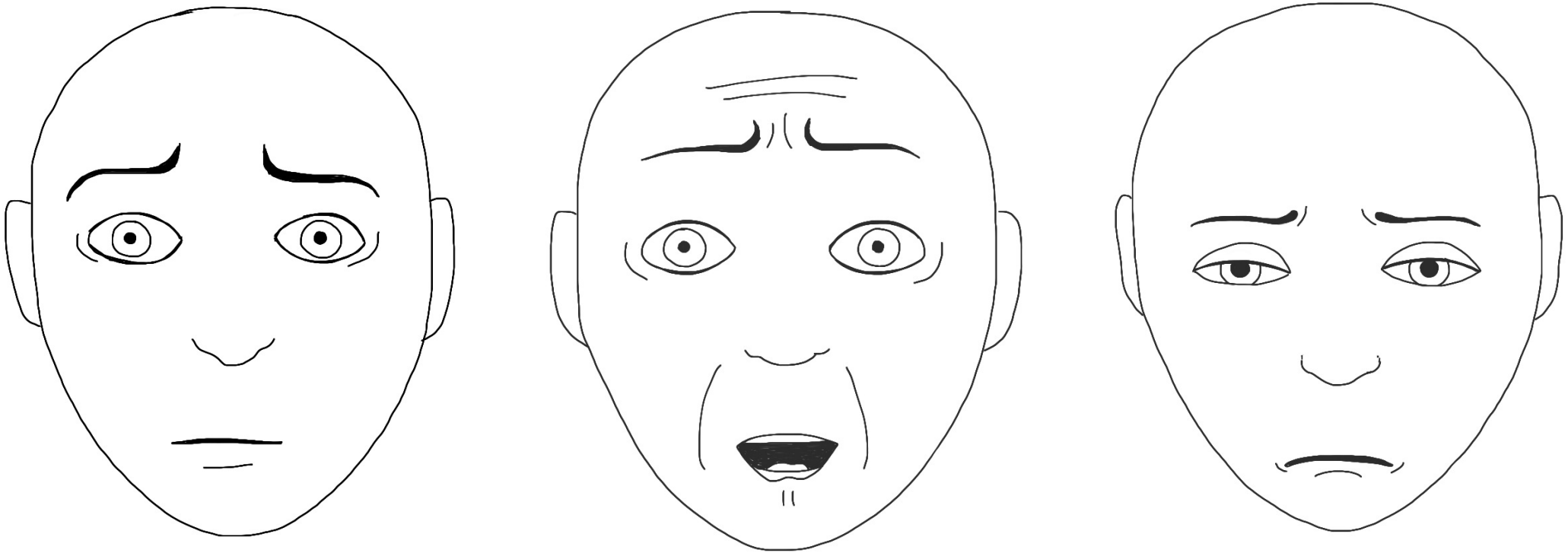
Benefits of using moral models that are interpretable enough for you to understand and correct them

1) You make the moral model more accurate

2) You become more clear about what <u>you</u> think is morally right and wrong, and why



You do the training and fine-tuning when you are calm and have time to think.

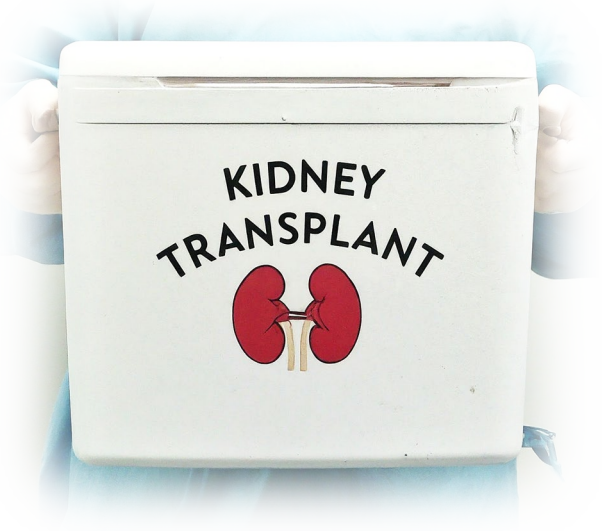


What your calm, rested, and thoughtful self would think is morally right.

Decisions that don't align well with the moral model you said you want to use.

We have been applying this approach to kidney allocation decisions, and are extending it to other contexts as well

Kidney transplants



Other scarce medical resources


First, surges in Covid-19 infections led to shortages of hospital beds and staff. Now it's oxygen

By Kristen Holmes and Aya Elamroussi, CNN

Updated 10:24 AM ET, Mon August 30, 2021

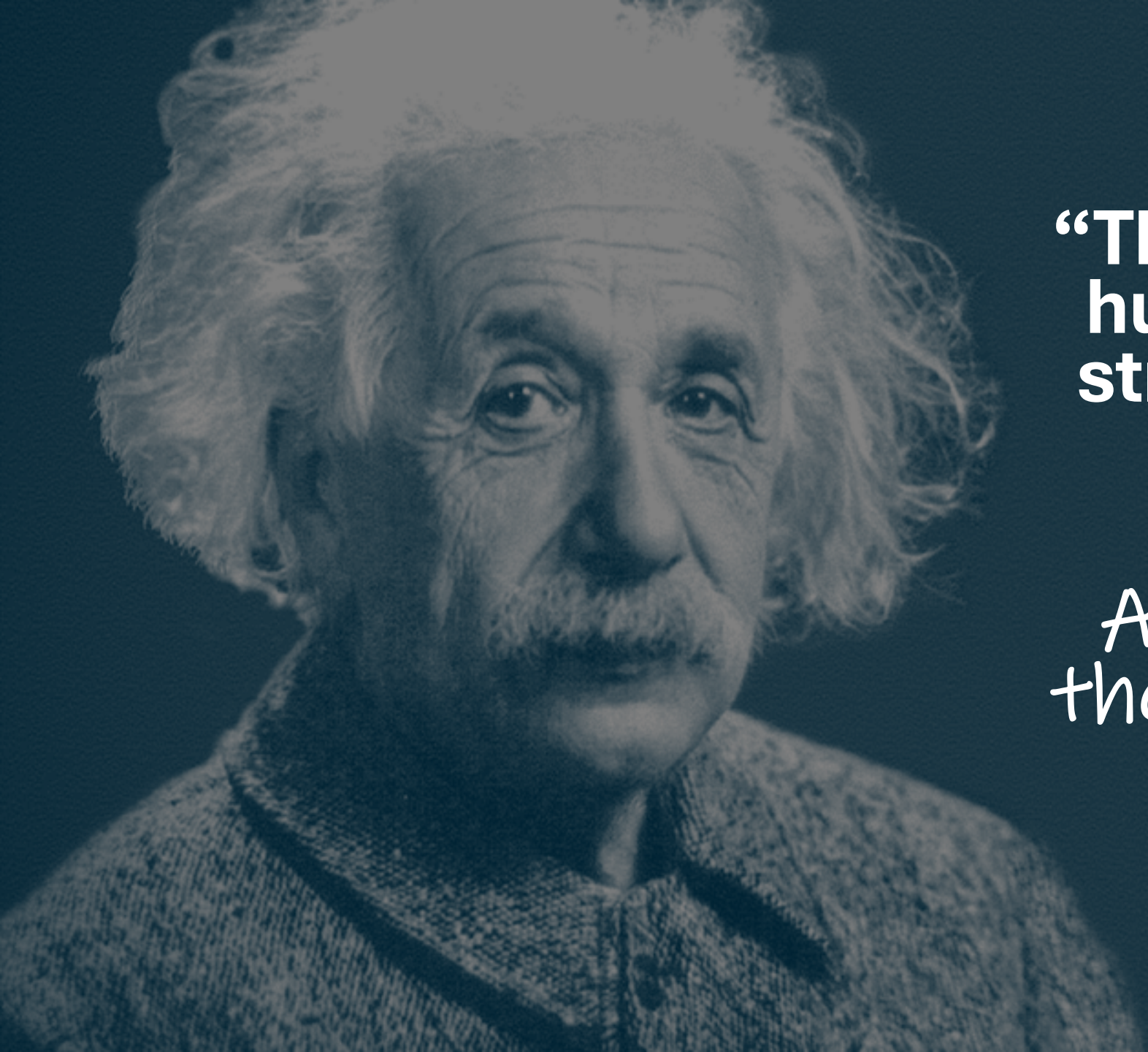
More from CNN

NBA legend Kareem Abdul-Jabbar calls for unvaccinated players to...

"It was an easy decision": McConnell announces he got a Covid-19...

Ad Hearing Aid Advice

Job applications and promotions



“The most important human endeavor is striving for morality in our actions.”

AI can be part of the solution, not just the challenge.