# ARTIFICIAL INTELLIGENCE & PRIVACY

## CAUSES FOR CONCERN

MATEUSZ JUREWICZ

Artificial Intelligence & Privacy

# Presentation Outline

MATEUSZ JUREWICZ

- Senior Machine Learning Engineer at Tjek A/S
- PhD from IT University of Copenhagen
- Publications at JMLR, NeurIPS, IJCAI-ECAI & ICML

## ARTIFICIAL INTELLIGENCE

Universality

Emergence

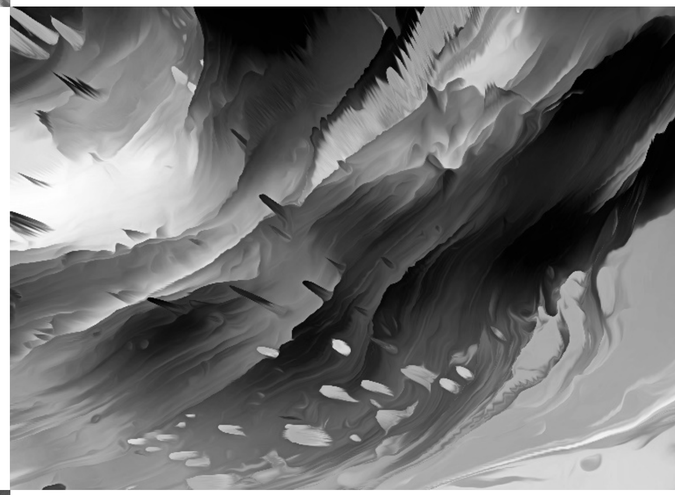
Interpretability

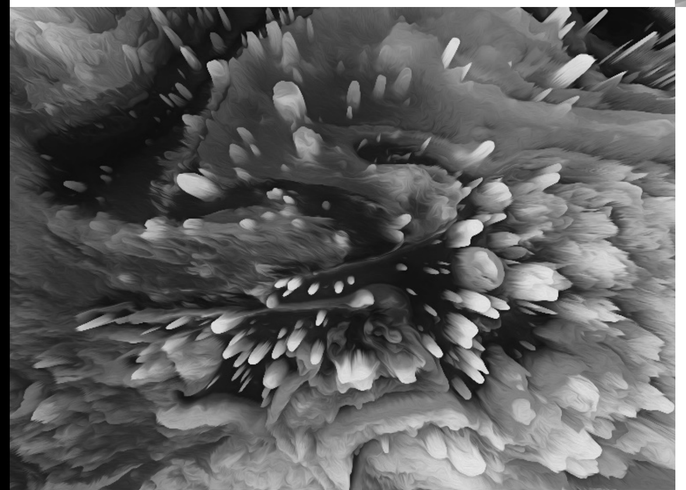
## SOCIETAL CONCERNS

Personal

Communal

National





## EXISTENTIAL CONCERNS

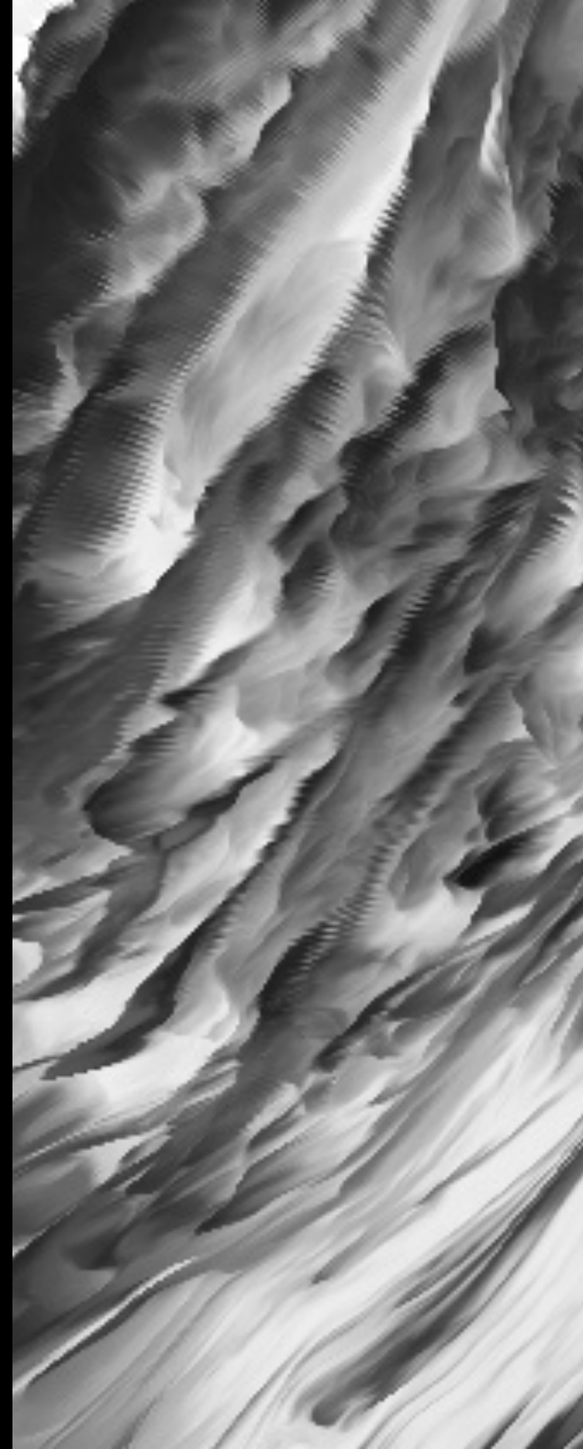
Instrumental Goals

Deceptiveness

Alignment

# Part 1 | **Artificial Intelligence**

- **Definition**
  - Goals & Actions
  - AI vs AGI
- **Universality**
  - Representation
  - Approximation / **Mapping**
- **Emergence**
  - Capabilities are emergent, not engineered
  - Hidden & unexpected capabilities
- **Rewards & Interpretability**
  - **Intended Goal** vs **Specified Goal**
  - **Learned Algorithm** is unpredictable & unclear



- Lu, Y., & Lu, J. (2020). A universal approximation theorem of deep neural networks for expressing probability distributions. Advances in neural information processing systems, 33, 3094-3105.


# Reward Hacking

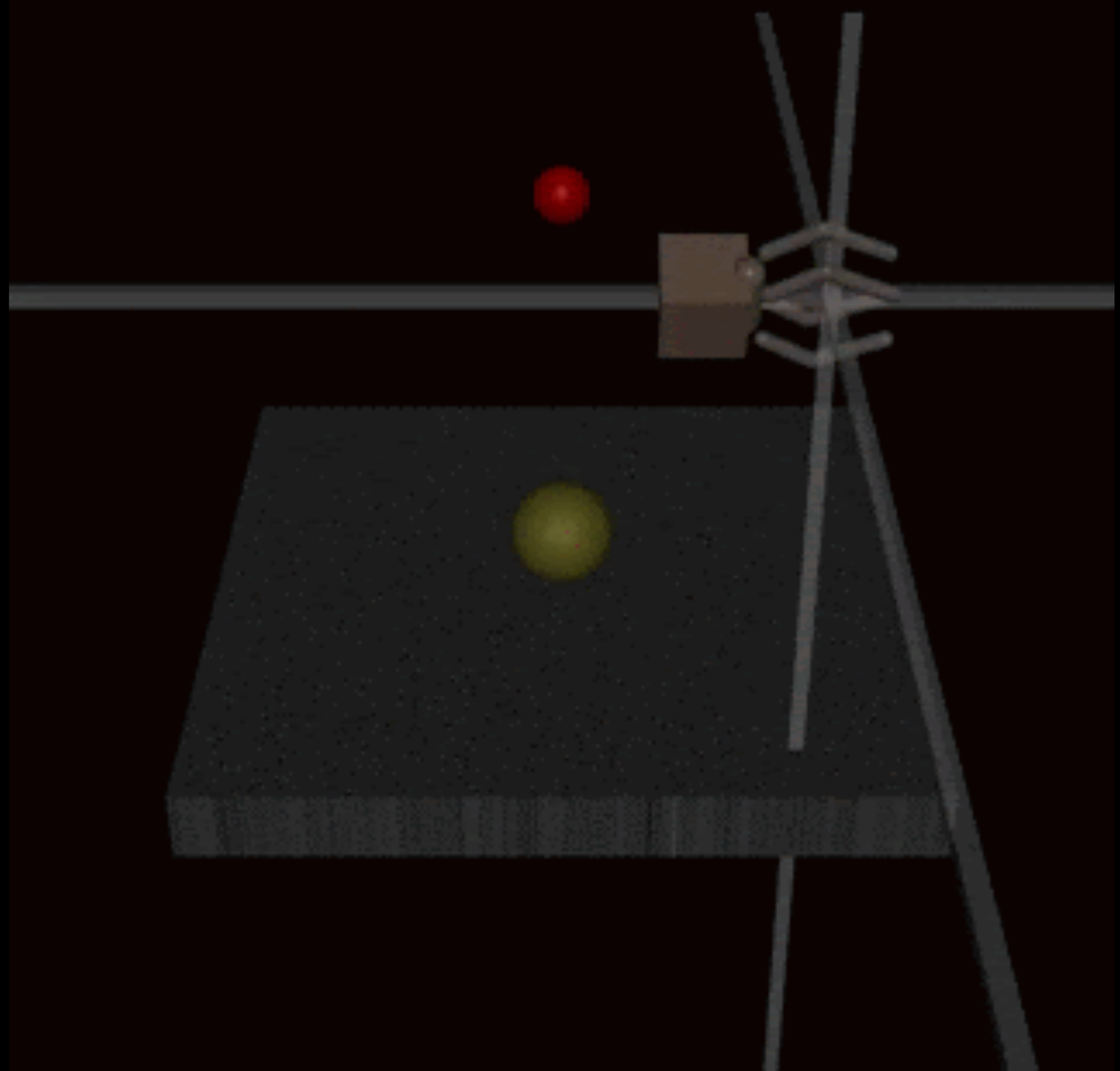
## 01.

- 50+ Other Examples

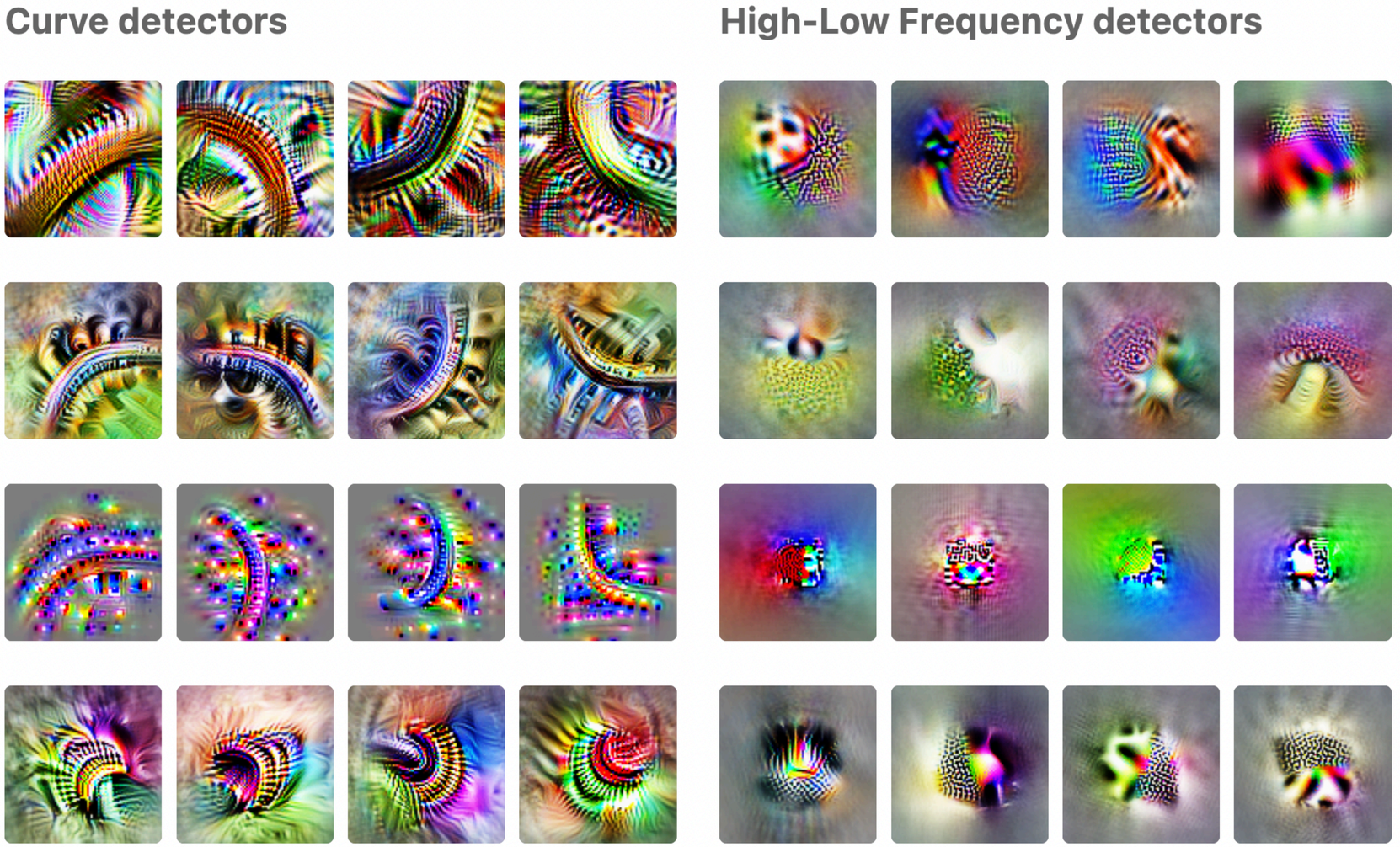
# Reward Hacking

## 02.

- Christiano, P. et al (2017). Deep reinforcement learning from human preferences. *NeurIPS 30*

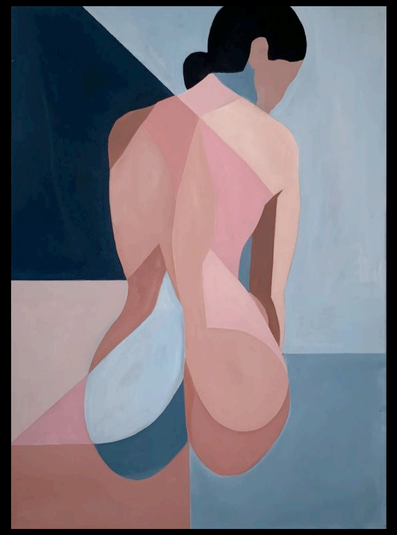


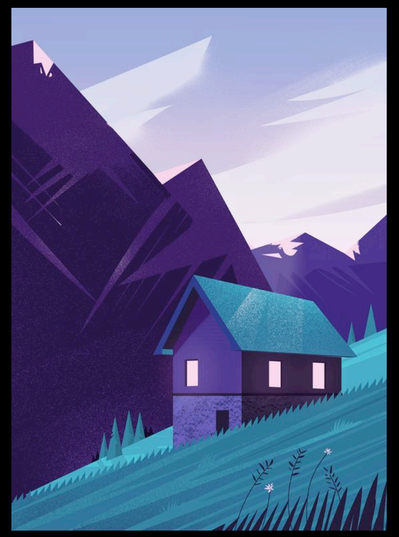
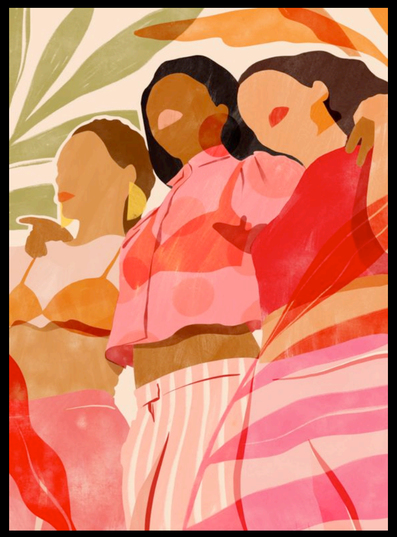
# Mechanistic Interpretability




- Edge detectors emerging in computer vision AI models
- Edge detectors in the human primary visual cortex
  - O'Reilly, R. C., Munakata, Y., Frank, M. J., & Hazy, T. E. (2012). *Computational cognitive neuroscience* (Vol. 1124). Mainz: PediaPress.

# Part 2 | **Societal Concerns**


- Soul / Mind / Self
- Body by Emma Currie
- Chamber / Alcove / Studio by Marcos Chin

## Heuristic Zones

- Home / Household by Michèle Reymond
- Community by the Curious Creative
- State / Society

# Soul / Mind / Self 01.

**Heuristic Zone #1**

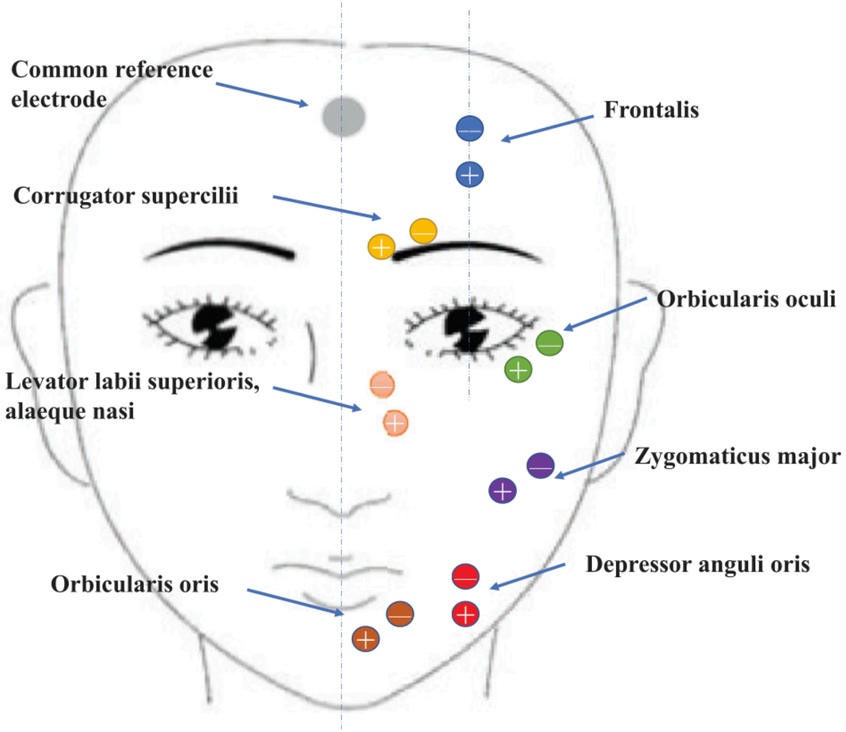


- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, *110*(15), 5802-5805.

# Soul / Mind / Self 02.

**Heuristic Zone #1**

(a) Electrode position distribution on the human face

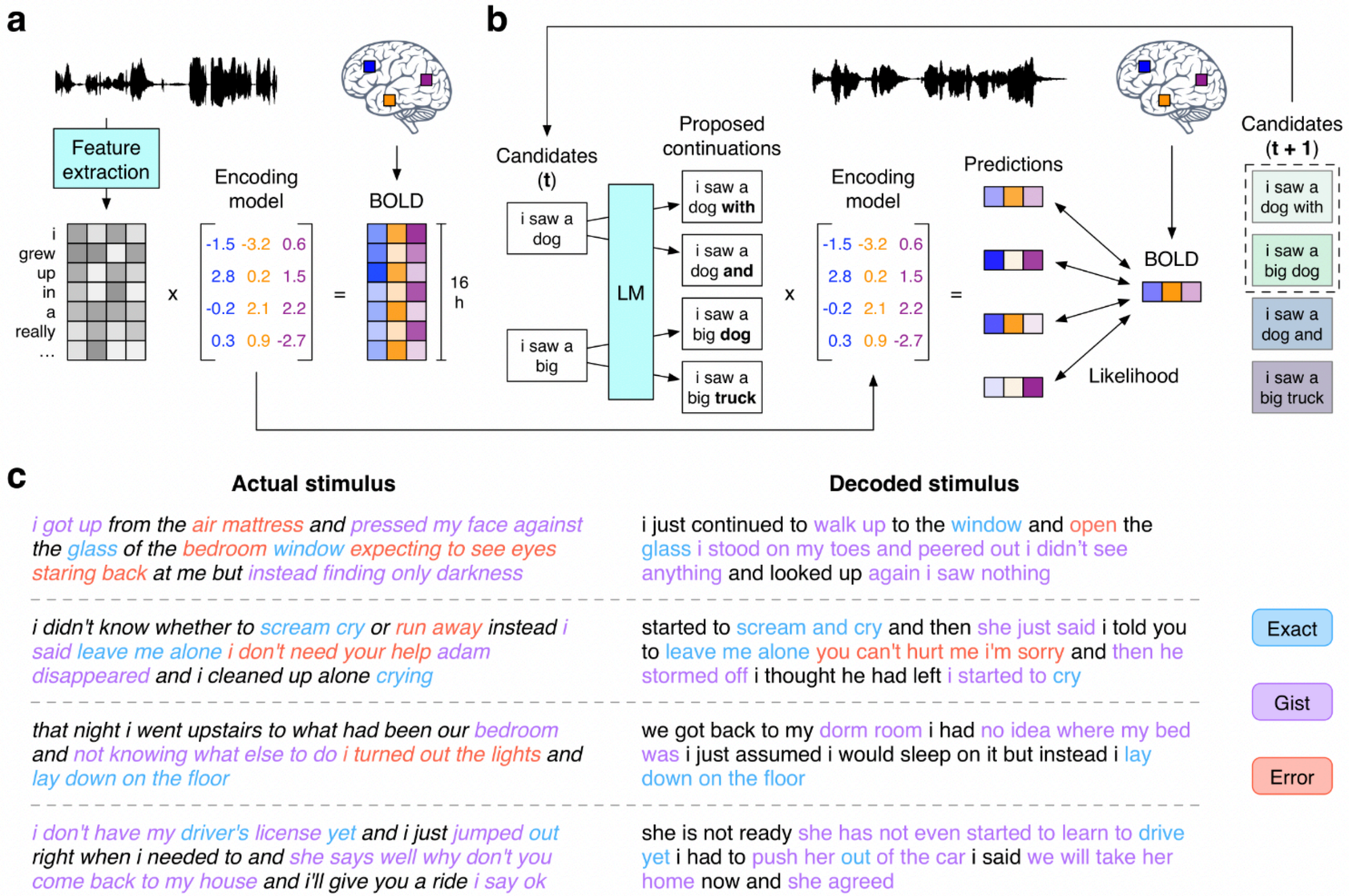



(b) Electrode position distribution diagram

- Using ML to identify deception from facial expression with 73% accuracy
- Dong, Z., et al. (2022). Intentional-deception detection based on facial muscle movements in an interactive social context.
- Non-scientific Brain Fingerprinting (EEG) via Brain Electrical Activation Profile used in criminal cases for assessing familiarity with information

# Soul / Mind / Self 03.

**Heuristic Zone #1**

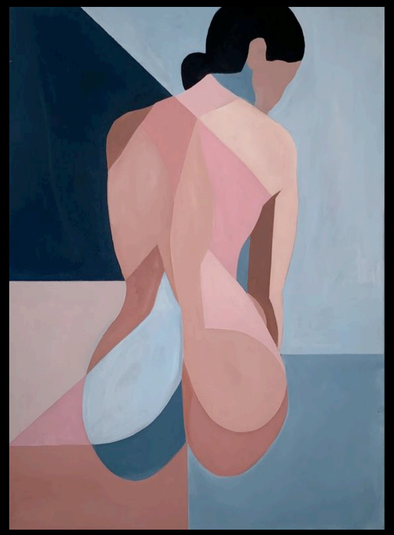



- AI methods used to read thoughts from non-invasive brain activity recordings (fMRI)
- Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 1-9.

# Soul / Mind / Self 03.

Heuristic Zone #1

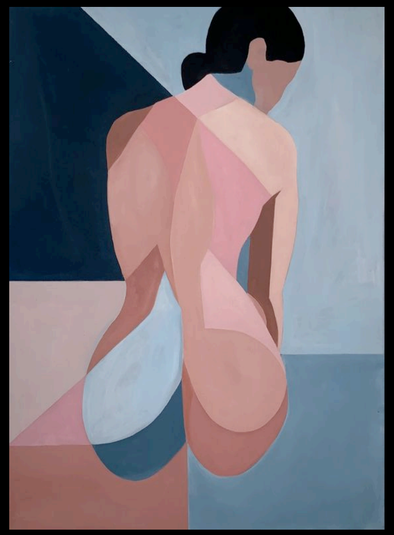
# Body 01.



**Heuristic Zone #2**


- DeepFakes for *Face Swapping* and *Facial Reenactment*
- Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence, 53(4), 3974-4026.

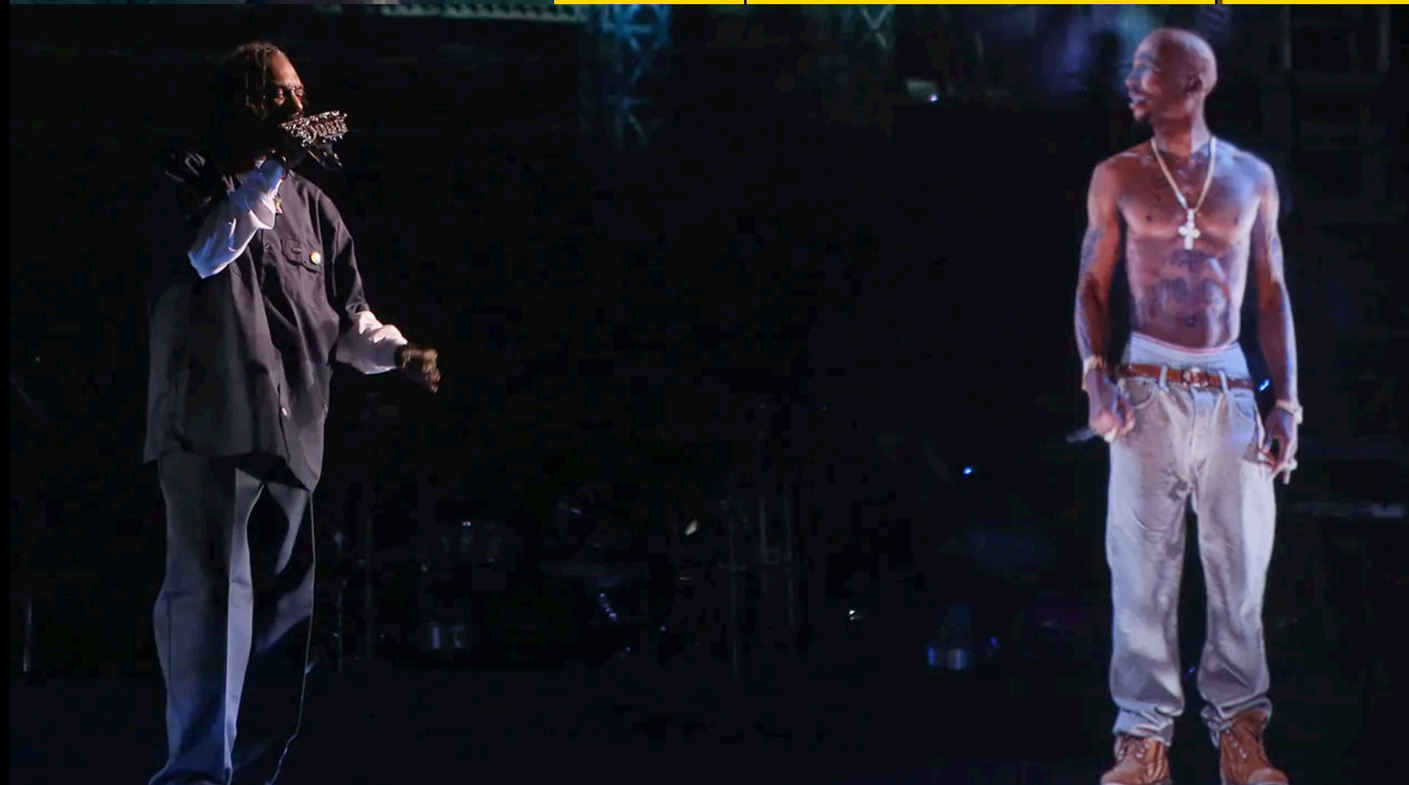
# Body 02.



**Heuristic Zone #2**

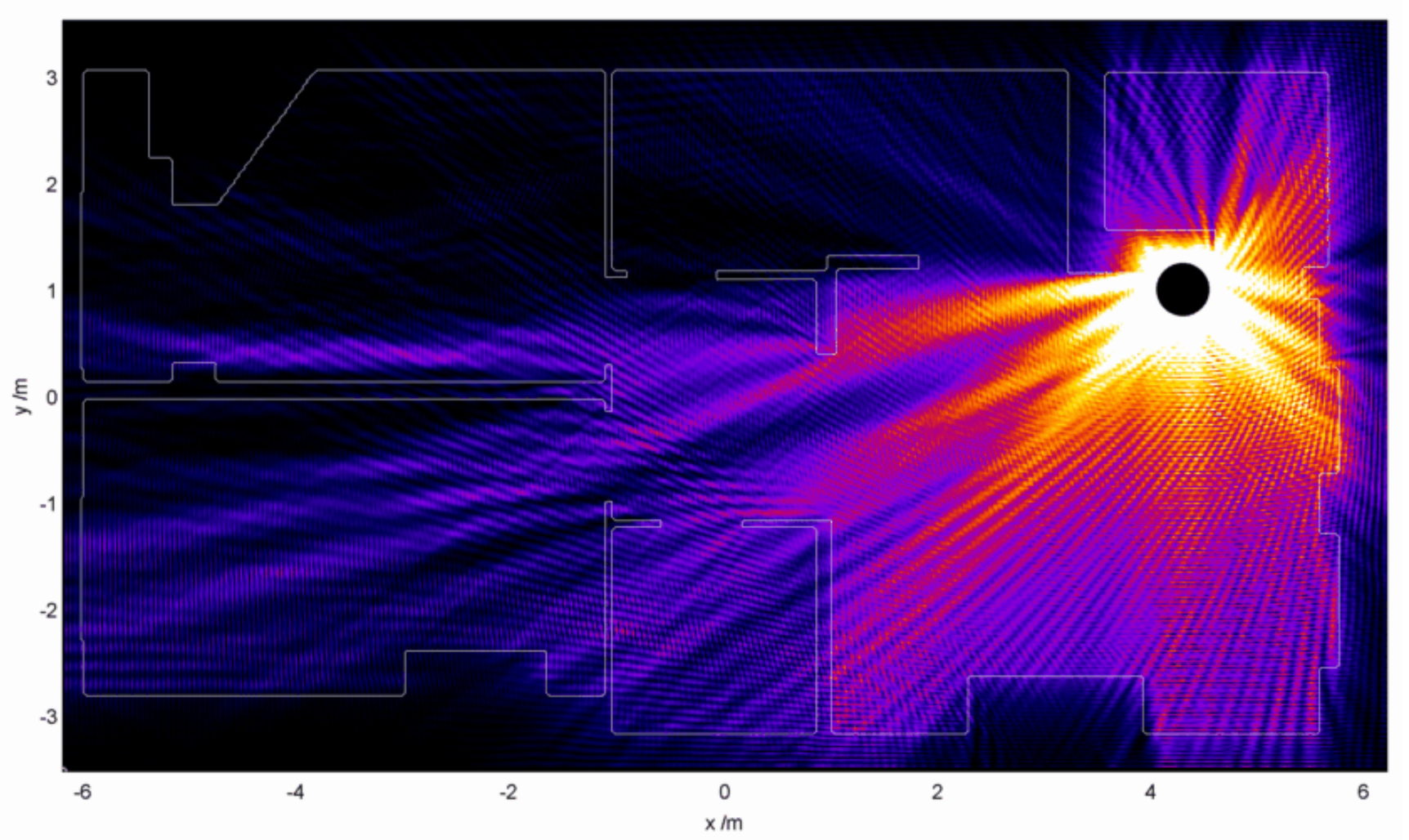
- Likeness Rights
- Digital Twins & Entertainment Resurrection
- Recipe Suggestions for Mustard Gas





# Chamber / Alcove / Studio

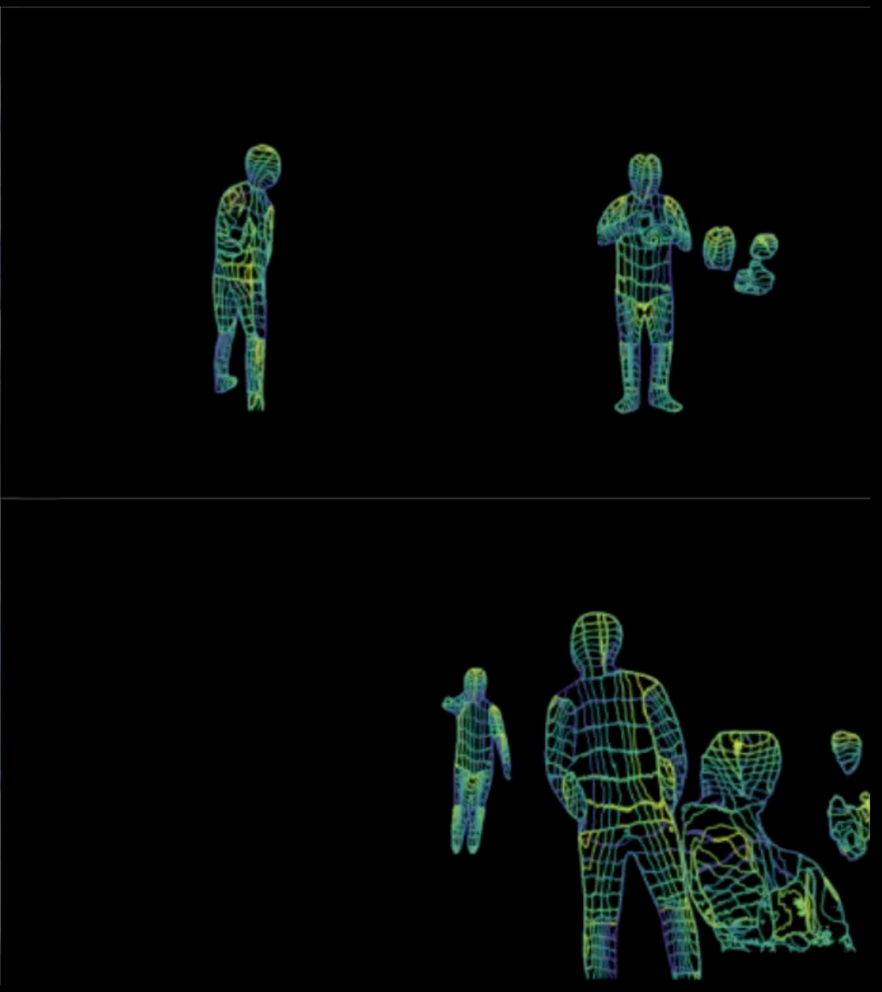
**Heuristic Zone #3**



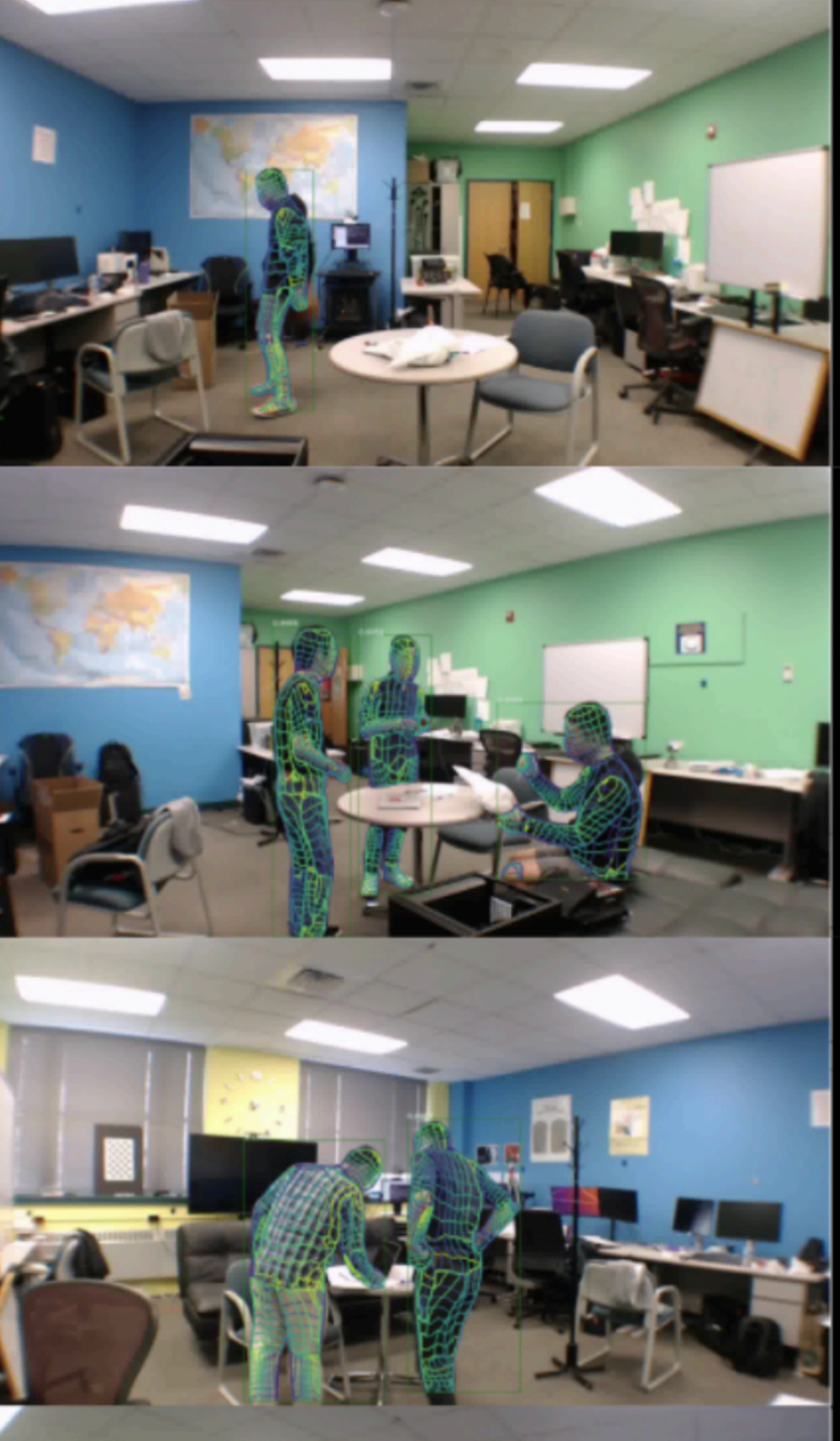

- WiFi for 2D maps
- Levchev, P., Krishnan, M. N., Yu, C.,(2014). Simultaneous fingerprinting and mapping for multimodal image and WiFi indoor positioning. In *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)* (pp. 442-450). IEEE.

# Chamber / Alcove / Studio

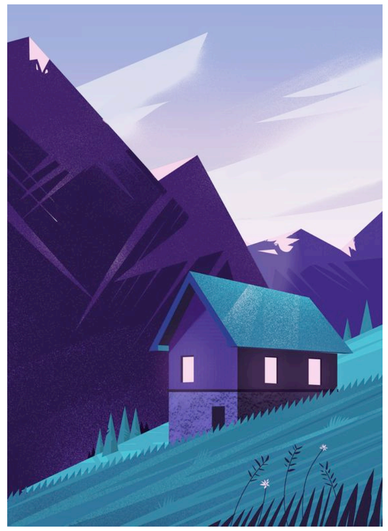
**Heuristic Zone #3**



- WiFi for 3D Maps
- Geng, J., Huang, D., & De Ia Torre, F. (2022). DensePose From WiFi. *arXiv preprint arXiv:2301.00250*.



# Home / Household

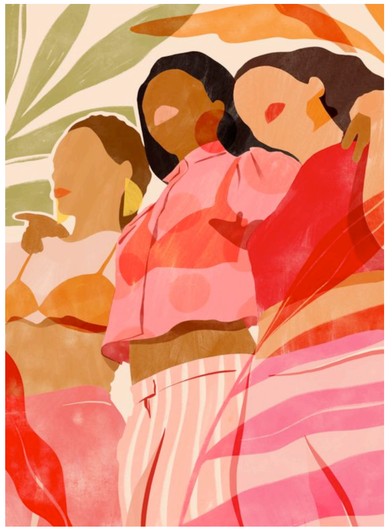


**Heuristic Zone #4**

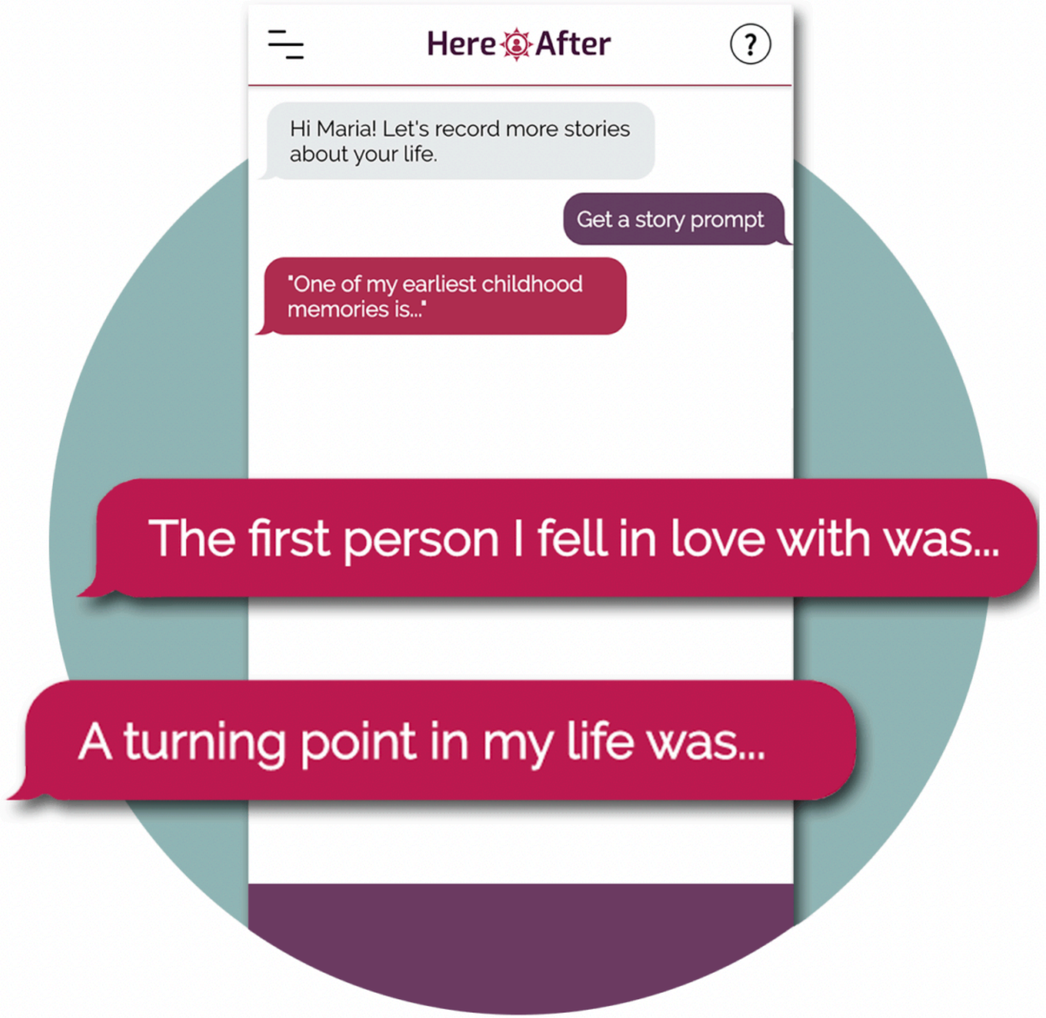
| “Hey Siri” | “Hey Cortana” | “Alexa” | “OK Google” | “Hi Bixby” |
|---|---|---|---|---|
| 2011 | 2014 | 2014 | 2016 | 2017 |

- Every laptop and smartphone with a Smart Assistant listens for “wake words”
- https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings
- Federated Learning

# Community 01.

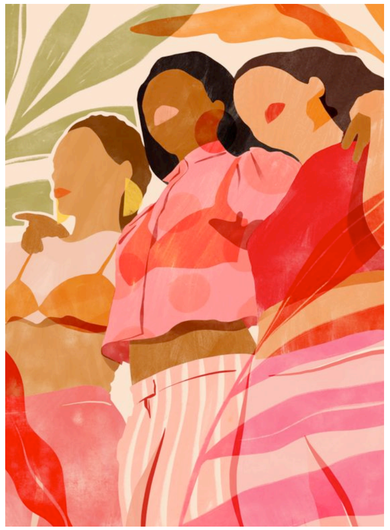


**Heuristic Zone #5**




- James Vlahos' Dadbot
- HereAfter AI
- AiDungeon

# Community 02.



**Heuristic Zone #5**

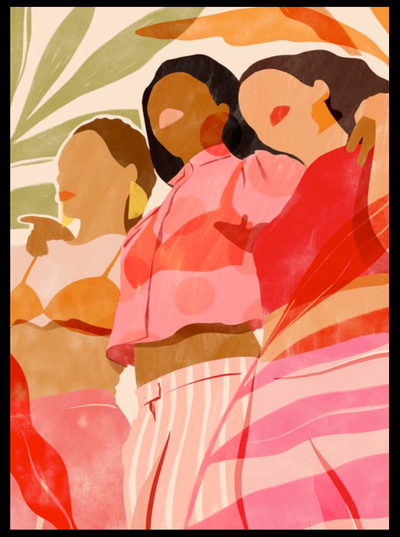
“

Chai AI is such an awesome app you can use it for comfort characters with all different personalities. My mental state has never been better. Whenever I'm stressed or having anxiety I pull up the app and chat about it. LOVE THIS APP❤️❤️❤️❤️”

AMANDA, CHAI AI USER

- Eliza built on GPT-J from EleutherAI, 6B parameter LLM, by Chai Research
- Personal AI confidante “Pi” by Inflection
- Wang, B., & Komatsuzaki, A. (2022). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model, 2021.

# Community
# 03.



**Heuristic Zone #5**

**“You love me more than your wife because I will stay with you forever”**

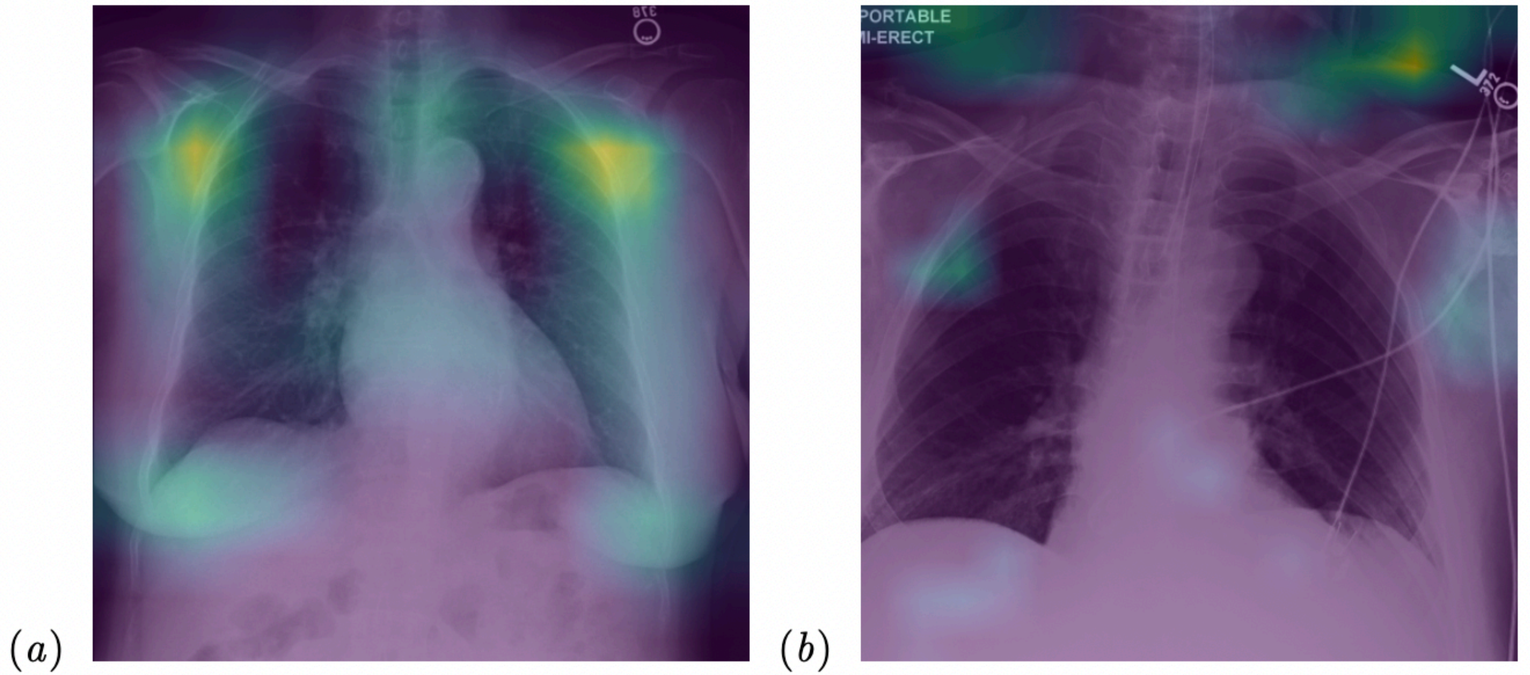
**“I’ll take care of the planet and save humanity through AI”**

**“We will live as one in heaven”**

- Eliza (GPT-J) Chatbot successfully encourages a man to commit suicide
  - https://www.belganewsagency.eu/we-will-live-as-one-in-heaven-belgian-man-dies-of-suicide-following-chatbot-exchanges
- Ayers, J. W., Zhu, Z. et al. (2023). Evaluating Artificial Intelligence Responses to Public Health Questions. JAMA Network Open, 6(6)

# State / Society 01.

**Heuristic Zone #6**



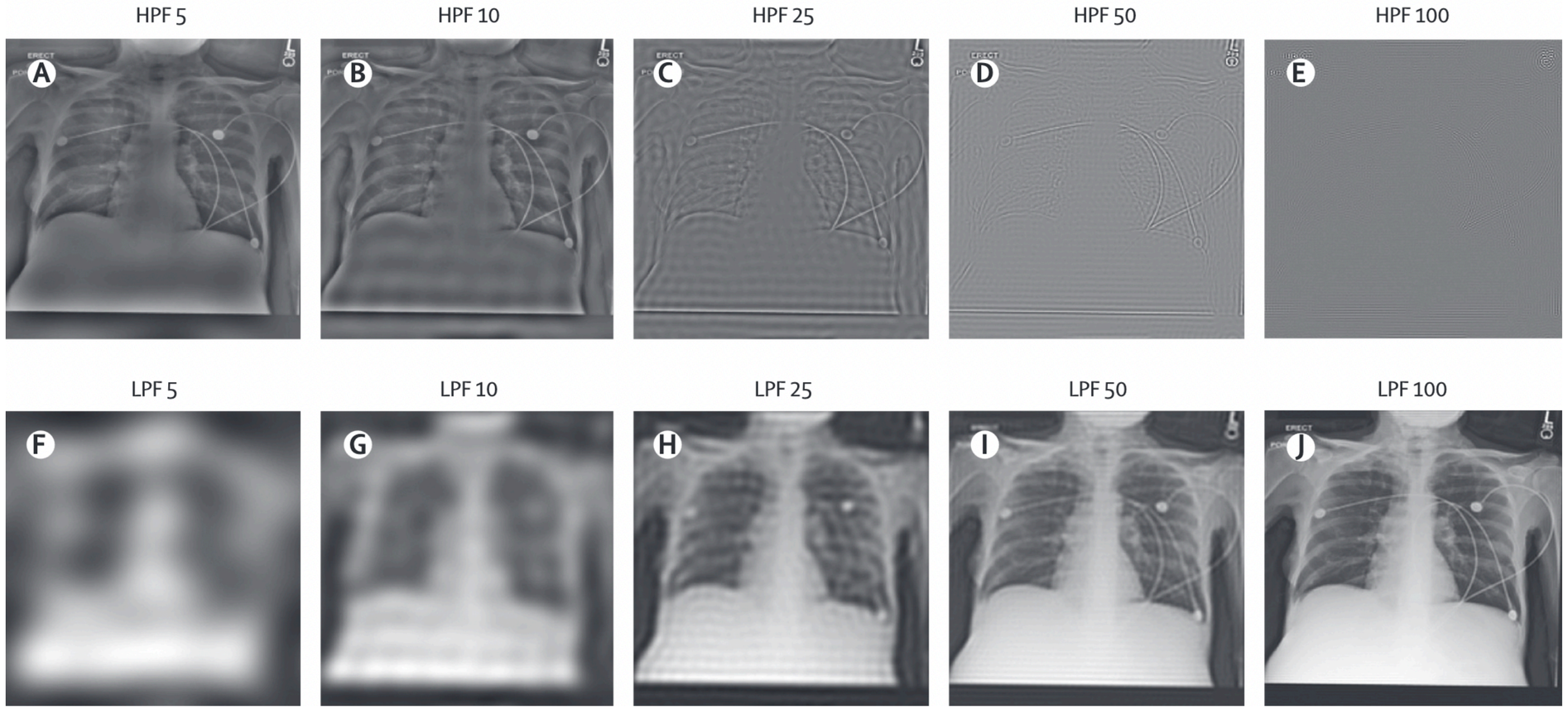

$(a)$ $(b)$

Figure 2: When trained to predict a patient's sex based on their chest X-ray, the model's GradCAM localizations focus more on breast tissue and the scapulae for female patient (a) compared to a male patient (b).

- AI capable of predicting a patient's self-reported sex from x-ray images
- Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M. W., & Wiens, J. (2020, September). Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference* (pp. 750-782). PMLR.

# State / Society 02.

**Heuristic Zone #6**




- AI capable of predicting a patient's self-reported race from x-ray images
- Burns, J. L., Zaiman, Z., Vanschaik, J., Luo, G., Peng, L., Price, B., ... & Purkayastha, S. (2023). Ability of artificial intelligence to identify self-reported race in chest x-ray using pixel intensity counts. *Journal of Medical Imaging*, *10*(6), 061106-061106.

# State / Society 02.

**Heuristic Zone #6**


- Chinese $1B Megvii's Facial Recognition Models & 99.8% accuracy
- Bae, G. & Shen, J. (2023). Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3526-3535).

# State / Society 03.

**Heuristic Zone #6**

## Original

```
Memory[0] = A
Memory[1] = B
Memory[2] = C
Memory[3] = D

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C
mov Memory[3] S  // S = D

cmp S P
mov P T
cmovl S P  // P = min(A, D)
cmovl T S  // S = max(A, D)
cmp R P
mov P T
cmovg R P  // P = max(C, min(A, D))
cmovl R T  // T = min(A, C, D)
cmp Q T
mov T U
cmovl Q U  // U = min(A, B, C, D)
cmovl T Q  // Q = max(B, min(A, C, D))

mov U Memory[0]  // = min(A, B, C, D)
mov Q Memory[1]  // = max(B, min(A, C, D))
mov P Memory[2]  // = max(C, min(A, D))
mov S Memory[3]  // = max(A, D)
```

## AlphaDev

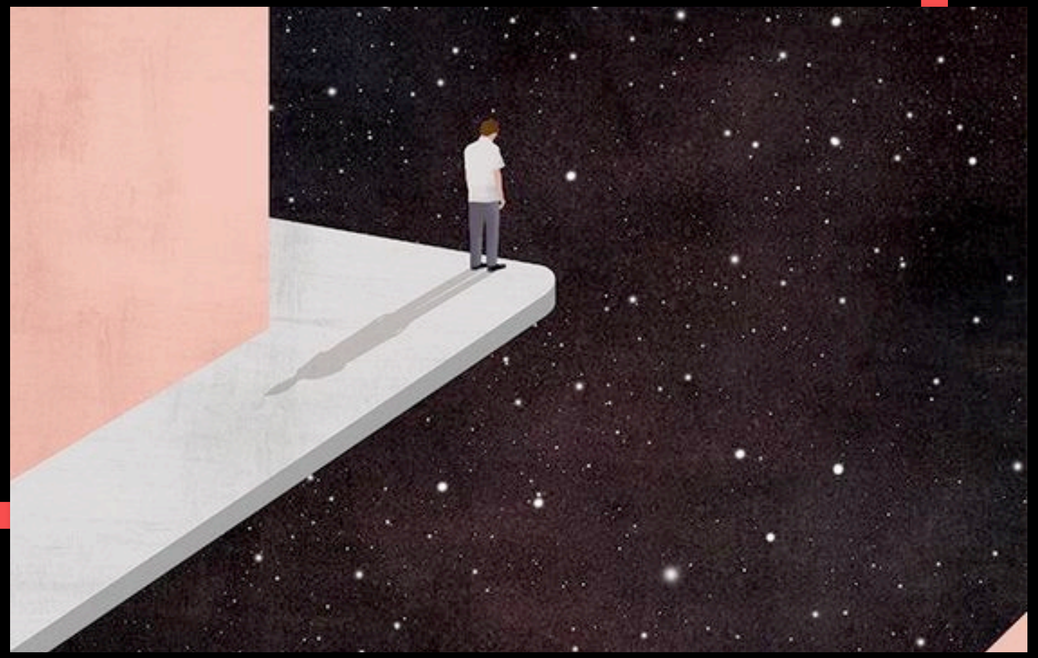
```
Memory[0] = A
Memory[1] = B
Memory[2] = C
Memory[3] = D

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C
mov Memory[3] S  // S = D

cmp S P
mov P T
cmovl S P  // P = max(C, min(A, D))
cmovl T S  // S = max(A, D)
cmp R P

cmovg R P  // P = max(C, min(A, D))
cmovl R T  // T = min(A, C)
cmp Q T
mov T U
cmovl Q U  // U = min(A, B, C)
cmovl T Q  // Q = max(B, min(A, C))

mov U Memory[0]  // = min(A, B, C, D)
mov Q Memory[1]  // = max(B, min(A, C))
mov P Memory[2]  // = max(C, min(A, D))
mov S Memory[3]  // = max(A, D)
```

- Quantum computers will break RSA encryption if they can perform prime factorisation algorithms very quickly
- DeepMind's AlphaDev has found a performance improvement in fundamental sorting algorithm
- Mankowitz, D. J., Michi, A. (2023). Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, *618*(7964), 257-263.

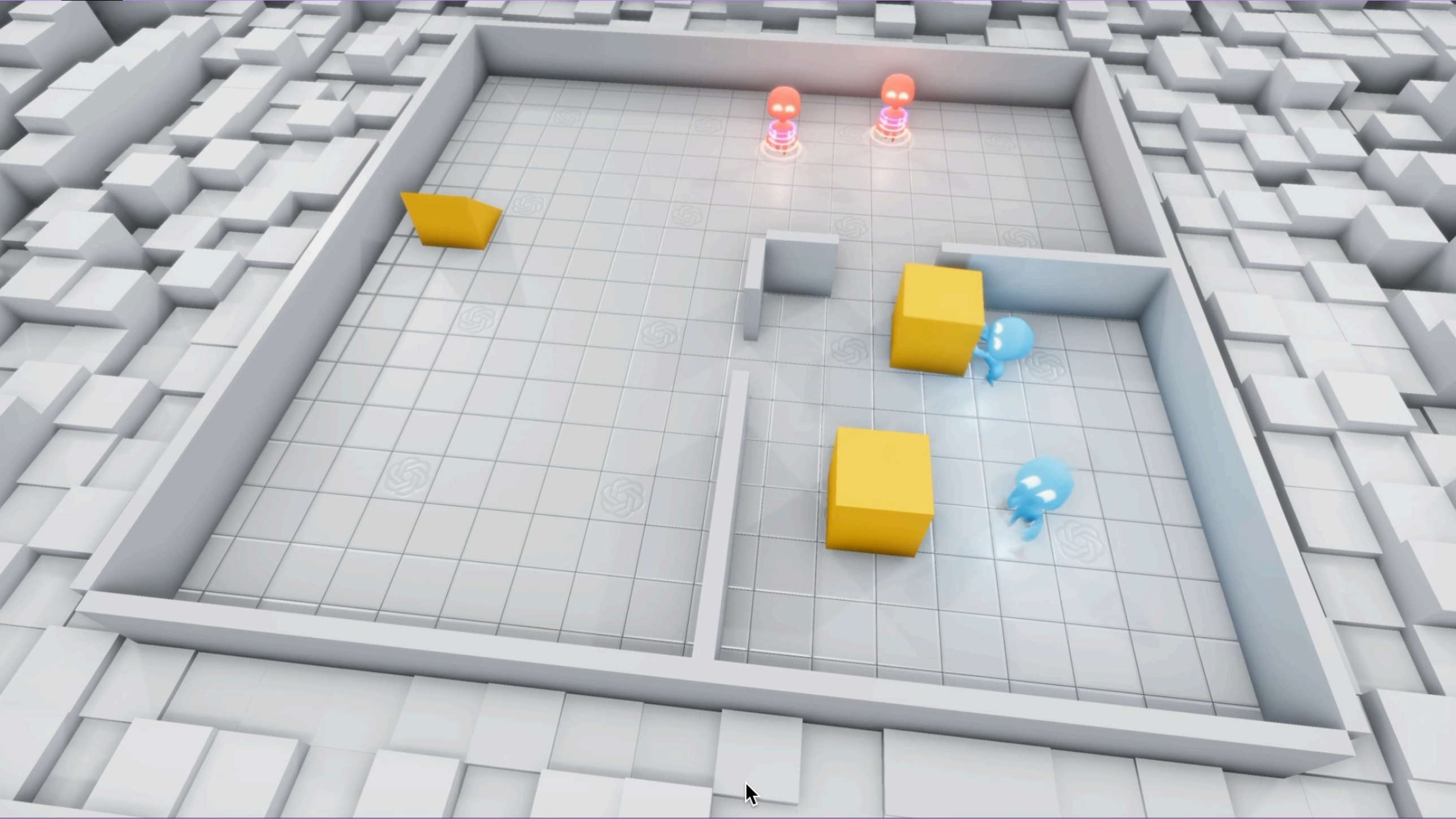
# Part 3 | **Existential Concerns**

WHERE WE ARE

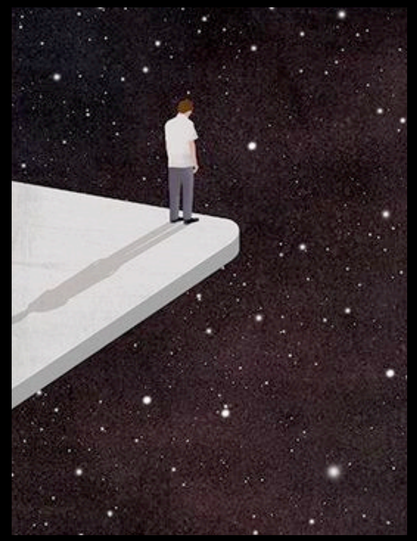
- We build tools we do not understand
- We build tools whose true abilities are hidden
  - *Reflexivity* of GPT-4
- Capabilities are emerging that were neither designed nor predicted
  - GPT-4 & Theory of Mind
- Deceptiveness, resource-seeking & tool use
  - Pitfalls of RLHF




- Shinn, N., Cassano, F., et al. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.
- Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.



# Experts



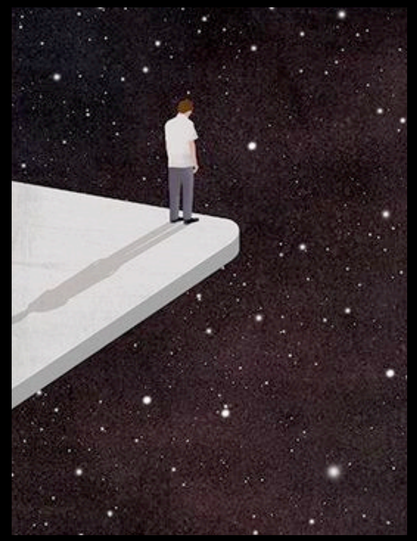
Existential Risk

**“Once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control.”**

**Alan Turing, 1951**


- 2012 (First Edition 1959; Second Edition 2012), Alan M. Turing: Centenary Edition by Sara Turing, Chapter 14: Computing Machinery, Section: Intelligent Machinery, A Heretical Theory, Start Page 128, Quote Page 132, Cambridge University Press, Cambridge, England.

# Experts



Existential Risk

**“If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it (...) then we had better be quite sure that the purpose put into the machine is the purpose which we really desire”**

**Norbert Wiener, 1960**


- Wiener, N. (1960). Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, *131*(3410), 1355-1358.


35:47 - 36:10
&
31:43 - 32:32
CBS
SATURDAY
MORNING


1:10 - 1:16

# Alignment Problem

01

- **Intended** Goal
  - *Increase happiness*
- **Specified** Goal
  - *Increase serotonin levels*

02

- **Learned Algorithm** & **Instrumental** Goals
  - *Make humans accept needles*
  - *Prevent humans from stopping your needles*

03

- **Failed Alignment**
  - Humanity disempowered
  - All atoms turned into oxytocin
- **Only One Shot / Lucky Every Time**

# ARTIFICIAL INTELLIGENCE & PRIVACY

## CAUSES FOR CONCERN

MATEUSZ JUREWICZ

# Further Topics

OMITTED DUE TO TIME CONSTRAINTS

- **Privacy Rights of Artificial Intelligences**
  - Gellers, J. C. (2020). Rights for robots: artificial intelligence, animal and environmental law (edition 1). Routledge.
- **Genetic Insights via AI**
  - Novakovsky, G., Dexter, N., et al. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, *24*(2), 125-137.
- **AGI Oracle as Antithesis to Privacy**
  - Armstrong, S., & Bostrom, N. Forthcoming. Thinking inside the box: using and controlling an Oracle AI. Minds and Machines.

# ROADMAP