Artificial Intelligence and Affective Computing

Jesse Hoey
University of Waterloo

CS886 Affective Computing 2024

@drjessehoey bayesact.ca jhoey@uwaterloo.ca

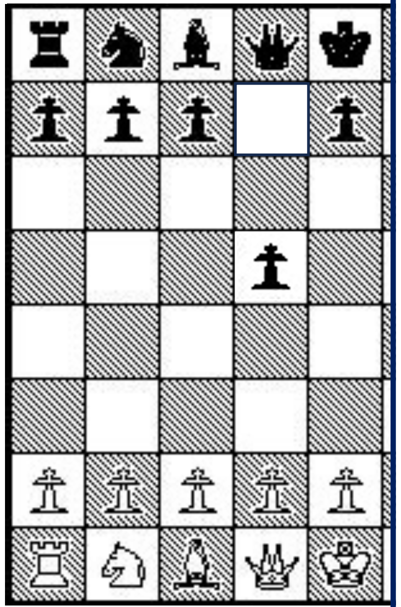
Artificial Intelligence: what is it?

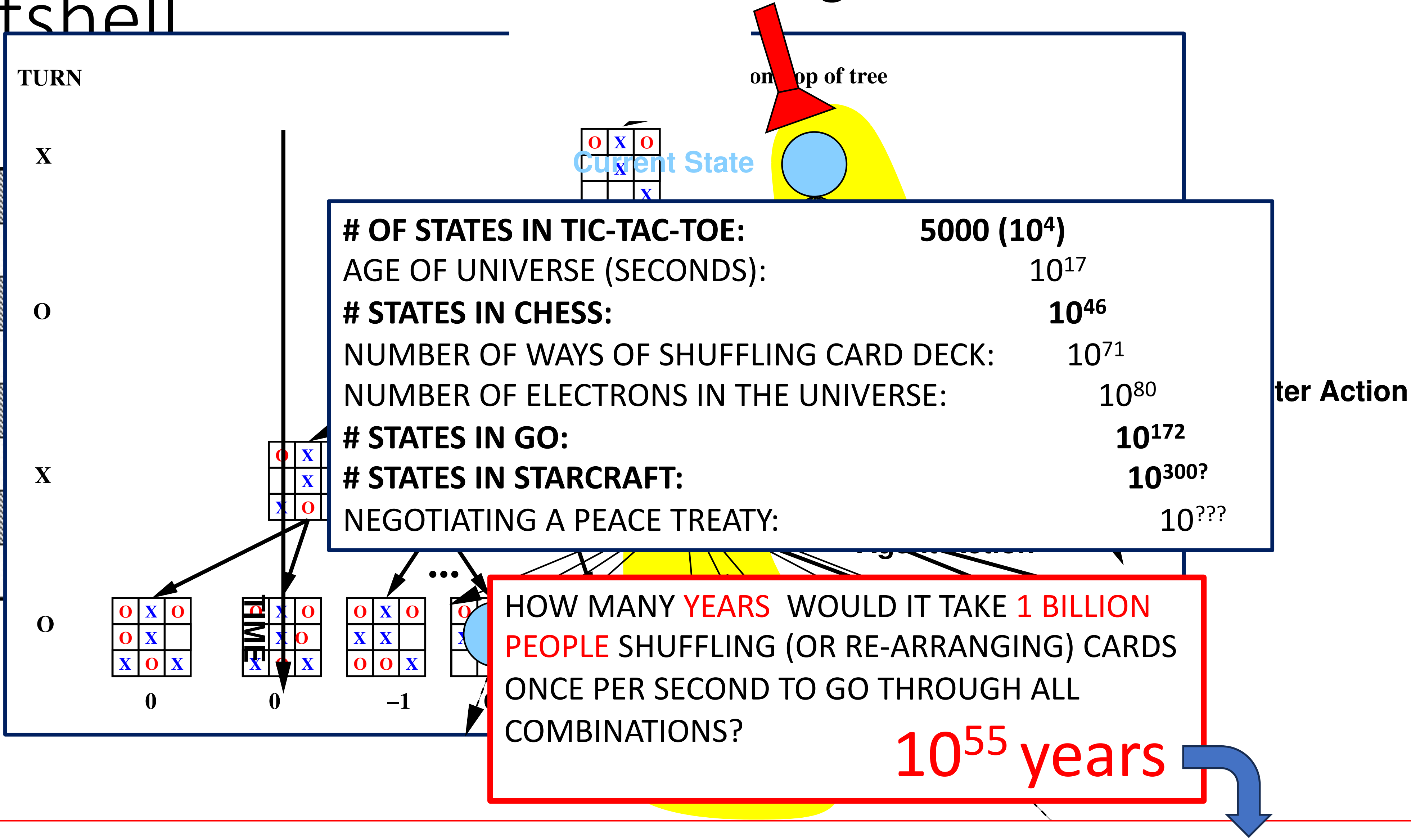
1996

2016

“A.I.” in a nutshell

heuristic ”flashlight”





# OF STATES IN TIC-TAC-TOE:	**5000 (10^4)**
AGE OF UNIVERSE (SECONDS):	10^{17}
# STATES IN CHESS:	**10^{46}**
NUMBER OF WAYS OF SHUFFLING CARD DECK:	10^{71}
NUMBER OF ELECTRONS IN THE UNIVERSE:	10^{80}
# STATES IN GO:	**10^{172}**
# STATES IN STARCRAFT:	**$10^{300?}$**
NEGOTIATING A PEACE TREATY:	$10^{???}$

HOW MANY YEARS WOULD IT TAKE 1 BILLION PEOPLE SHUFFLING (OR RE-ARRANGING) CARDS ONCE PER SECOND TO GO THROUGH ALL COMBINATIONS?

10^{55} years

10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

age of universe in years → 4,000,000,000

Turing Test

- Alan Turing: 1912-1954
- **Can machines think?** - the imitation game
- *The interrogator is connected to one person and one machine via a terminal, therefore can't see her counterparts. Her task is to find out which of the two candidates is the machine, and which is the human only by asking them questions. If the machine can "fool" the interrogator, it is intelligent.*

ELIZA (Joseph Weizenbaum, 1966)

- ME: Hi
- ELIZA: How do you do? What brings you to see me?
- ME: I’m thinking of taking a course in artificial intelligence.
- ELIZA: Why do you say that?
- ME: I’m worried about it
- ELIZA: Are you worried often?
- ME: when I take AI courses
- ELIZA: What happened when you take ai courses?
- ME: I usually spend long nights doing assignments
- ELIZA: Earlier you said you felt worried?
- ME: about AI courses

Loebner Prize

- CONFEDERATE: Hello, my name is Sam and I am the human
- JUDGE: Hi Sam, tell me which is larger, a TV screen or Belgium.
- CONFEDERATE: Belgium is a little bit larger.
- JUDGE: Wljso lwjejdo wleje elwjeo wjeol, isn’t it?
- CONFEDERATE: Is that Polish?
- ...
- JUDGE: Which is further away from this apartment, Queens or the Queen of England?
- CONFEDERATE: If the Queen is visiting Manhattan, she is closer.
- JUDGE: If the Queen is visiting Manhattan, won’t she want to abdicate?
- CONFEDERATE: I would.
- JUDGE: You would rather not be Queen, then, right?
- CONFEDERATE: I’d like to be Queen, but England’s damp. How about you?

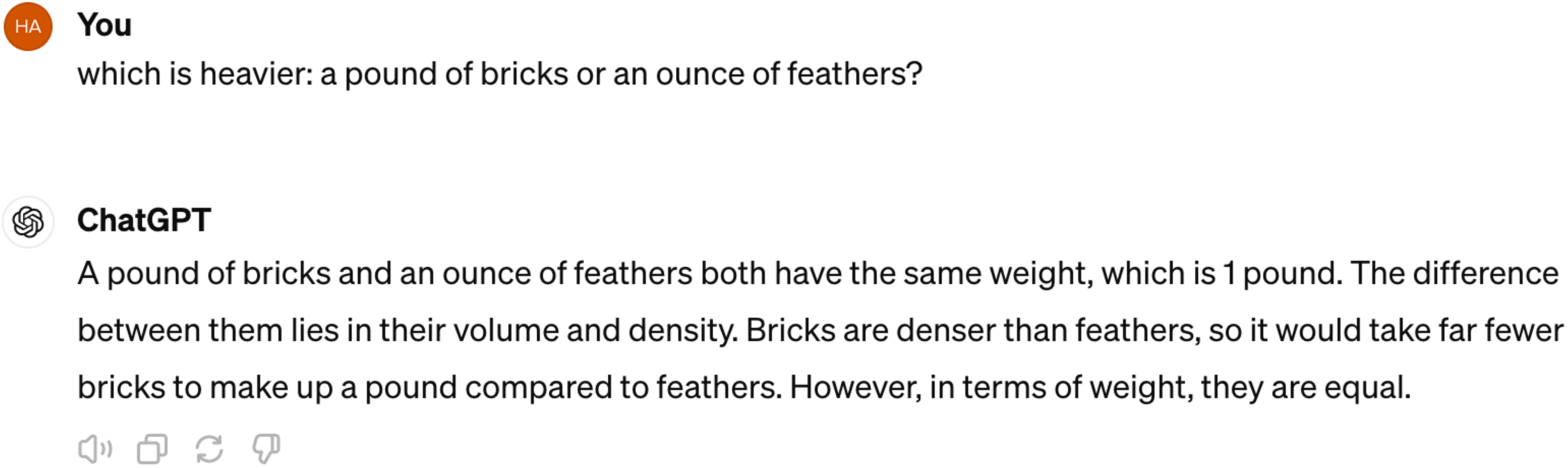
JabberWacky (Rollo Carpenter)

- JABBERWACKY: Hello, my name is George and I am the human.
- JUDGE: If Isjkel Iwehlx wouldn't you say?
- JABBERWACKY: No, he's still good.
- ...
- JUDGE: Which is further away from us, Queens or Inner Mongolia?
- JABBERWACKY: HA! HA! That was stupid. Tell me a GOOD joke this time.
- ...
- JUDGE: Which is larger, your brain or Belgium?
- JABBERWACKY: Oh the one on the far left without a doubt

BULLSHIT:
Indifferent to the truth of its outputs.

- Frankfurt "On Bullshit", 2005

ChatGPT (and the like):



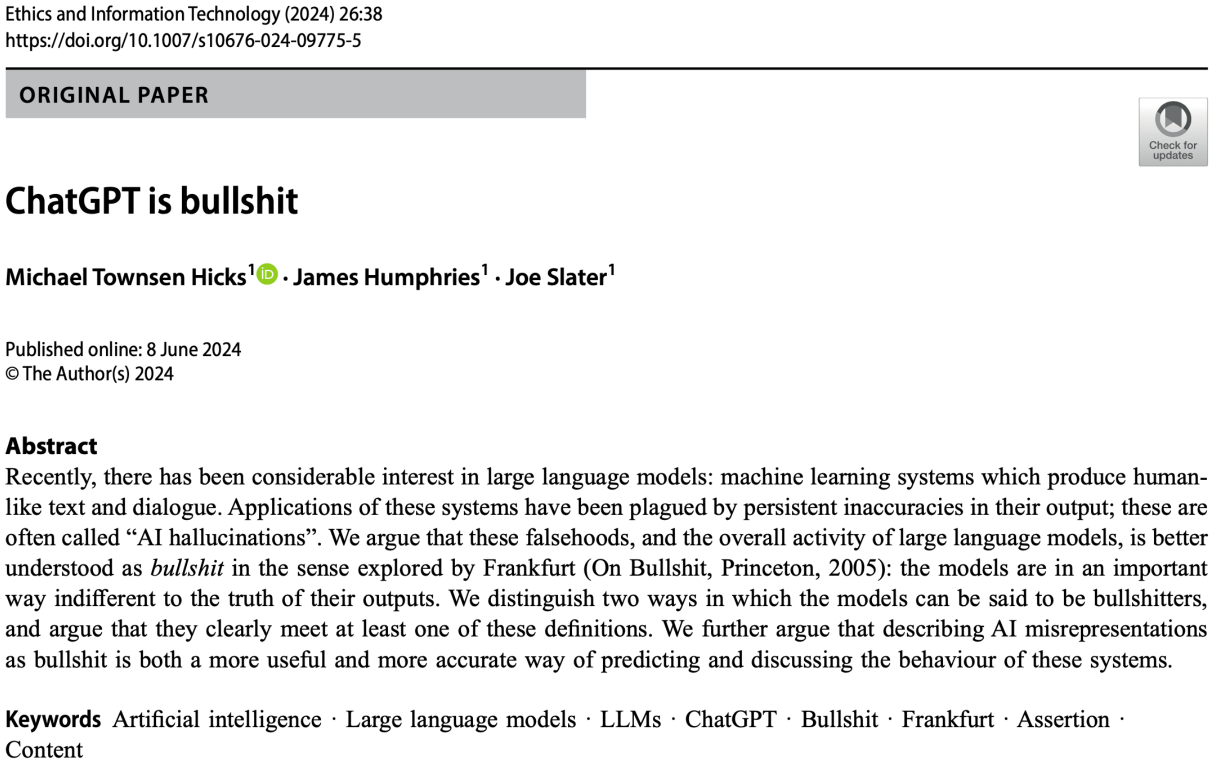
You

which is heavier: a pound of bricks or an ounce of feathers?

ChatGPT

A pound of bricks and an ounce of feathers both have the same weight, which is 1 pound. The difference between them lies in their volume and density. Bricks are denser than feathers, so it would take far fewer bricks to make up a pound compared to feathers. However, in terms of weight, they are equal.

ChatGPT is Jabberwacky on S---something



Ethics and Information Technology (2024) 26:38
https://doi.org/10.1007/s10676-024-09775-5

ORIGINAL PAPER

ChatGPT is bullshit

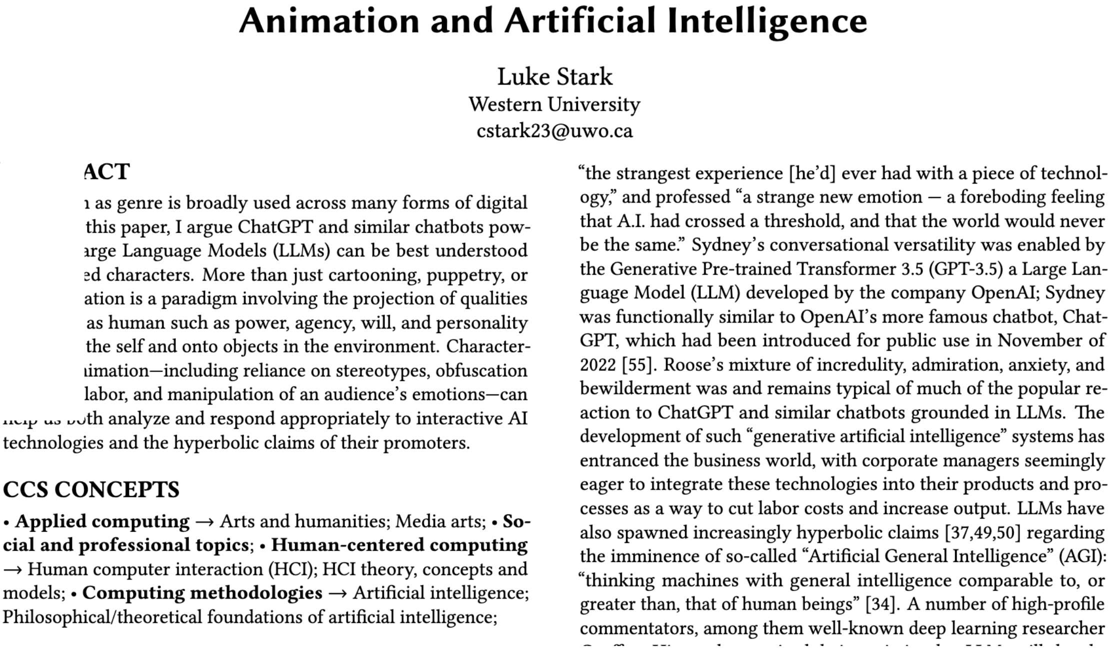
Michael Townsen Hicks[1] · **James Humphries**[1] · **Joe Slater**[1]

Published online: 8 June 2024
© The Author(s) 2024

Abstract
Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called “AI hallucinations”. We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

Keywords Artificial intelligence · Large language models · LLMs · ChatGPT · Bullshit · Frankfurt · Assertion · Content

FaCCT 2024



Animation and Artificial Intelligence

Luke Stark
Western University
cstark23@uwo.ca

ACT

ı as genre is broadly used across many forms of digital this paper, I argue ChatGPT and similar chatbots pow-arge Language Models (LLMs) can be best understood ed characters. More than just cartooning, puppetry, or ation is a paradigm involving the projection of qualities as human such as power, agency, will, and personality the self and onto objects in the environment. Character-imation—including reliance on stereotypes, obfuscation labor, and manipulation of an audience’s emotions—can oth analyze and respond appropriately to interactive AI technologies and the hyperbolic claims of their promoters.

CCS CONCEPTS

• **Applied computing** → Arts and humanities; Media arts; • **Social and professional topics**; • **Human-centered computing** → Human computer interaction (HCI); HCI theory, concepts and models; • **Computing methodologies** → Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence;

“the strangest experience [he’d] ever had with a piece of technology,” and professed “a strange new emotion — a foreboding feeling that A.I. had crossed a threshold, and that the world would never be the same.” Sydney’s conversational versatility was enabled by the Generative Pre-trained Transformer 3.5 (GPT-3.5) a Large Language Model (LLM) developed by the company OpenAI; Sydney was functionally similar to OpenAI’s more famous chatbot, ChatGPT, which had been introduced for public use in November of 2022 [55]. Roose’s mixture of incredulity, admiration, anxiety, and bewilderment was and remains typical of much of the popular reaction to ChatGPT and similar chatbots grounded in LLMs. The development of such “generative artificial intelligence” systems has entranced the business world, with corporate managers seemingly eager to integrate these technologies into their products and processes as a way to cut labor costs and increase output. LLMs have also spawned increasingly hyperbolic claims [37,49,50] regarding the imminence of so-called “Artificial General Intelligence” (AGI): “thinking machines with general intelligence comparable to, or greater than, that of human beings” [34]. A number of high-profile commentators, among them well-known deep learning researcher

Robotics

Shakey 1971

NOT bullshit

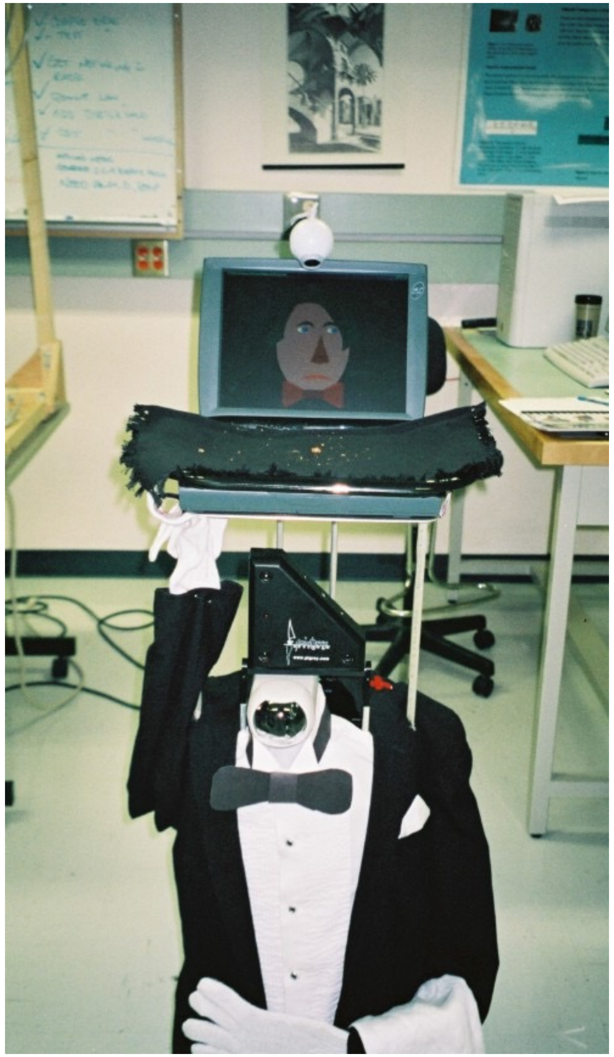
Robocup 2017

Emotional Robots?

BULLSHIT: Indifferent to the truth of its outputs

José the Robot

- 1999
- Winner of the “Hors D’oeuvres, Anyone?” AAAI Mobile robot competition in Seattle, WA.
- GREAT FAME ensued…

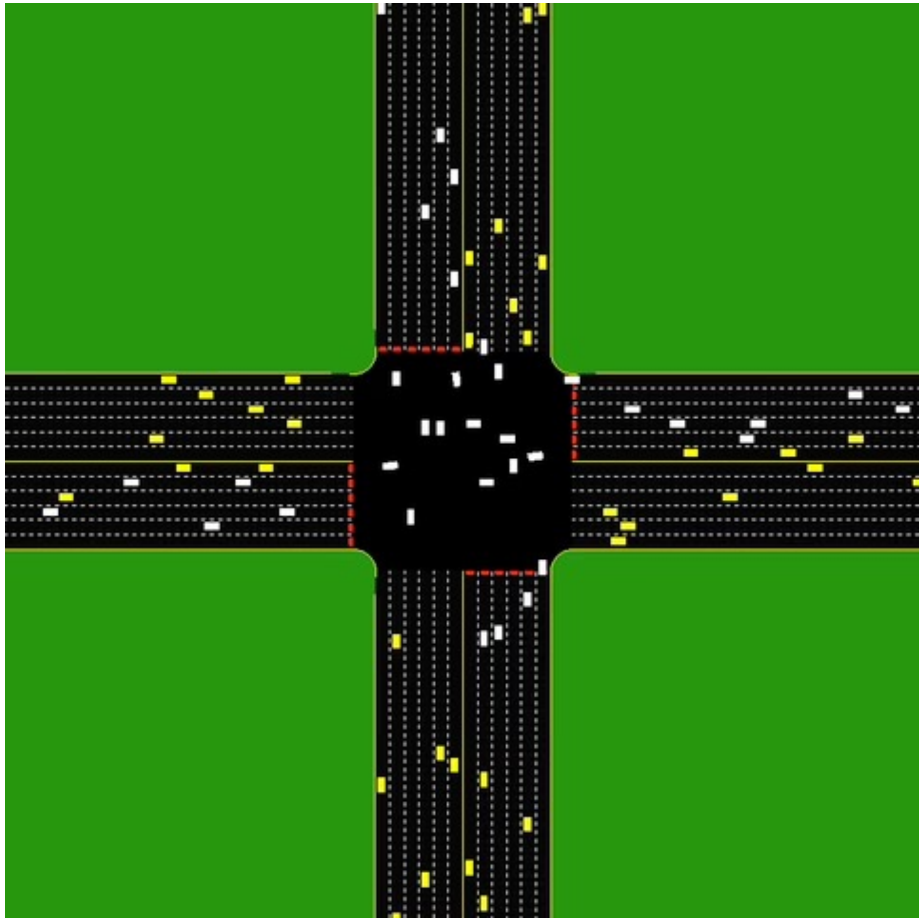


Autonomous Cars

STANLEY 2005: Darpa Grand Challenge

Sebastian Thrun Stanford

Peter Stone UT Austin



What’s missing: Human alignment

Are you going? Or should I go?
You go first.
What if I point a lot and flail my arms around?
This is confusing.
Wait, maybe you should go.
Let's just sit here and reflect.
What is this bullshit?

theoatmeal.com/blog/google_self_driving_car

“[these cars] have to learn to be aggressive in the right amount, and the right amount depends on the culture.”

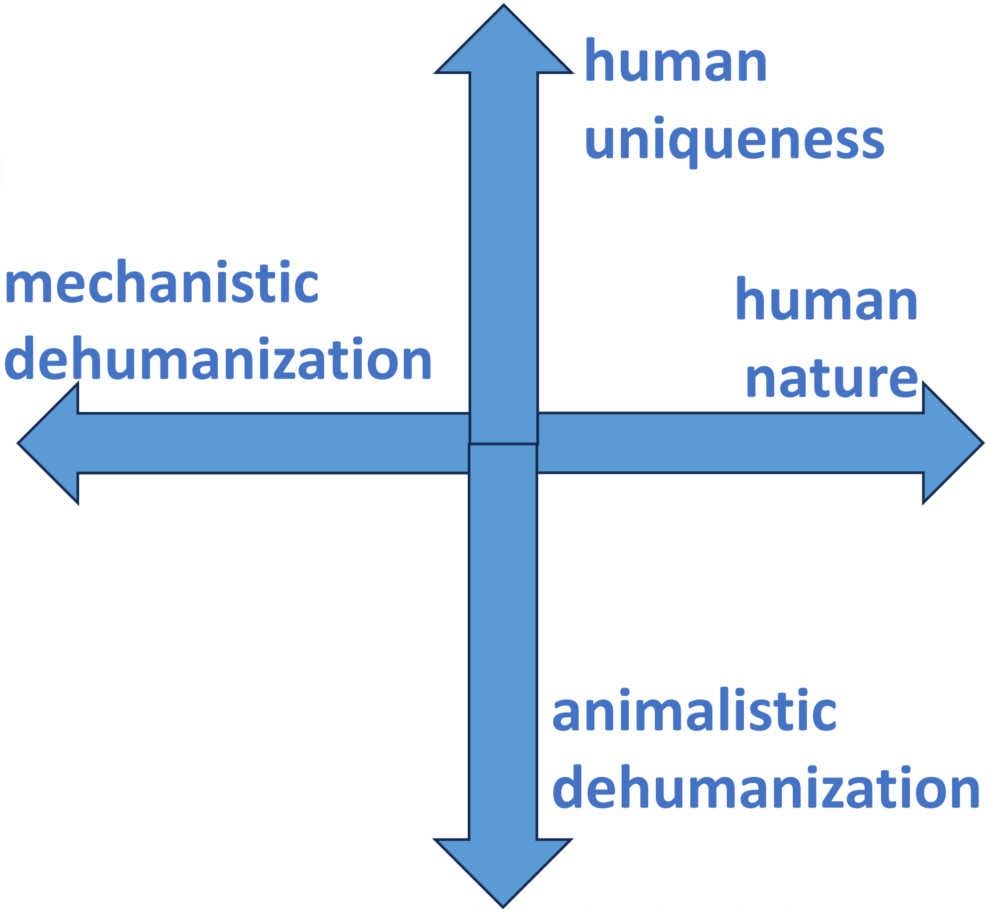
– Donald Norman, Design Lab, UCSD

from: New York Times “Googles Driverless Cars Run Into Problem: Cars With Drivers”, 02/09/2015.

Being Human

“A.I.”

rational, mature



human uniqueness

humans

mechanistic dehumanization

human nature

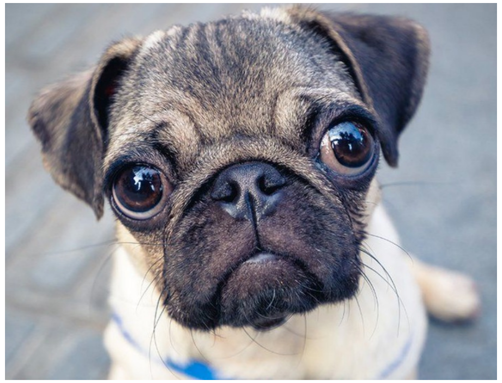
inertness, coldness

emotionality, warmth

machines

animalistic dehumanization

irrational, childlike



animals

adapted from: Nick Haslam, Stephen Loghnan, Yoshihisa Kashima and Paul Bain (2009). Attributing and Denying humanness to others. *European Review of Social Psychology*, 19:1, 55-85

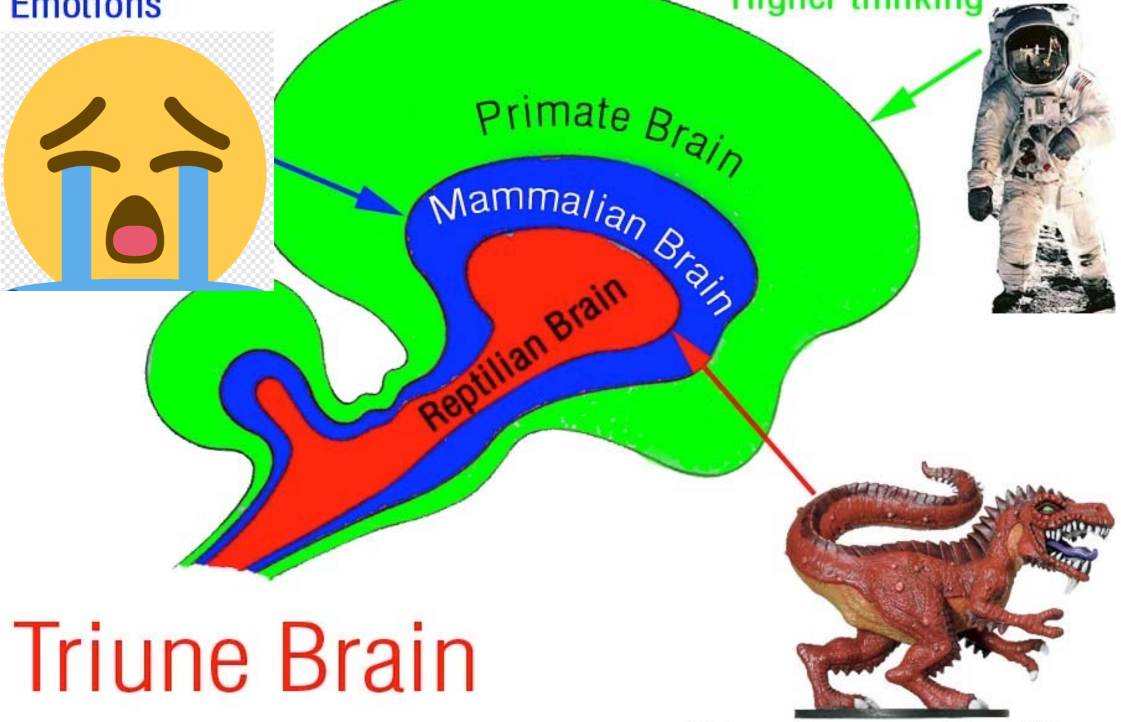
Phrenology 1800s …. Triune Brain 1960s …

FIRMNESS
VENERATION
BENEVOLENCE
SELF ESTEEM
CONSCIENTIOUSNESS
HOPE
SPIRITUALITY
IMITATION
SUAVITY
HUMAN NATURE
CONTINUITY
APPROBATIVENESS
SUBLIMITY
IDEALITY
CAUSALITY
COMPARISON
MIRTHFULNESS
INHABITIVENESS
CAUTIOUSNESS
FRIENDSHIP
SECRETIVENESS
ACQUISITIVENESS
CONSTRUCTIVENESS
TIME
MEMORY
LOCALITY
TUNE
COMBATIVENESS
DESTRUCTIVENESS
ALIMENTIVENESS
WEIGHT
COLOR
ORDER
SIZE
FORM
CALCULATION
LANGUAGE
PHILOPROGENITIVENESS
CONJUGAL LOVE
VITATIVENESS
AMATIVENESS

Phrenological Chart of the Faculties.

Intermediate: paleopallium
Limbic System
Emotions

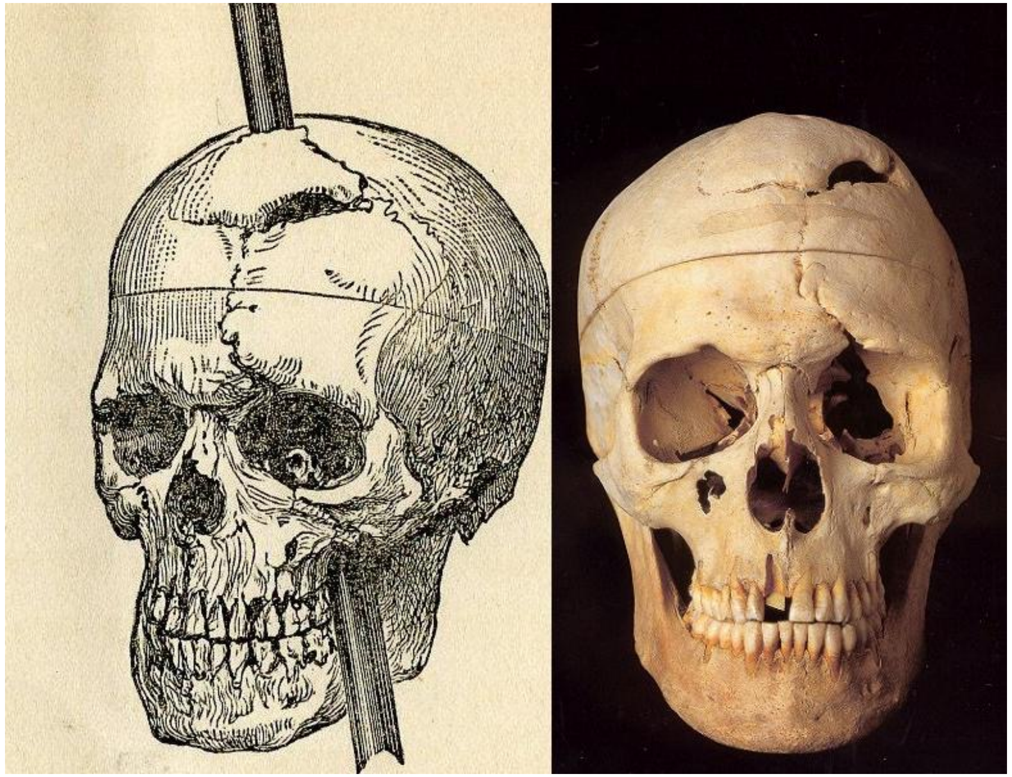
Rational Brain
Neocortex: neopallium
Higher thinking



Primative: archipallium
Survival, aggression

What are emotions?

- Antonio Damasio *Descarte's Error 1994*
- *Phineas Gage 1823-1860*
- Suffered brain damage (frontal lobe)
- Was perfectly good at reasoning and language
- Made disastrous decisions, or could not make decisions
- lacked "somatic markers" - "gut feelings" about decisions



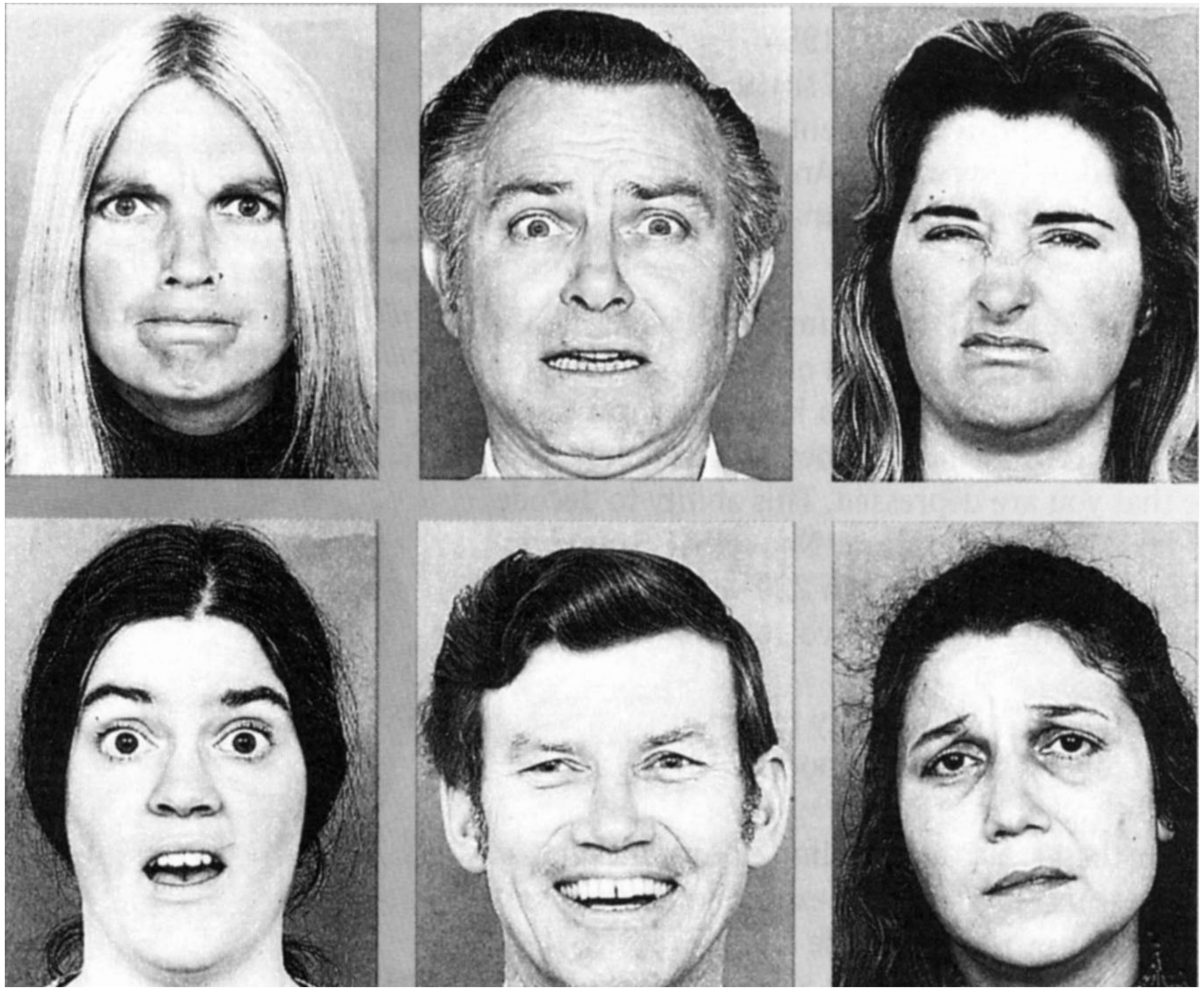
Affective Computing

Roz Picard 1997 book “Affective Computing”:
This book proposes that we give computers the ability to recognize, express and in some cases “have” emotions. Is this not absurd?

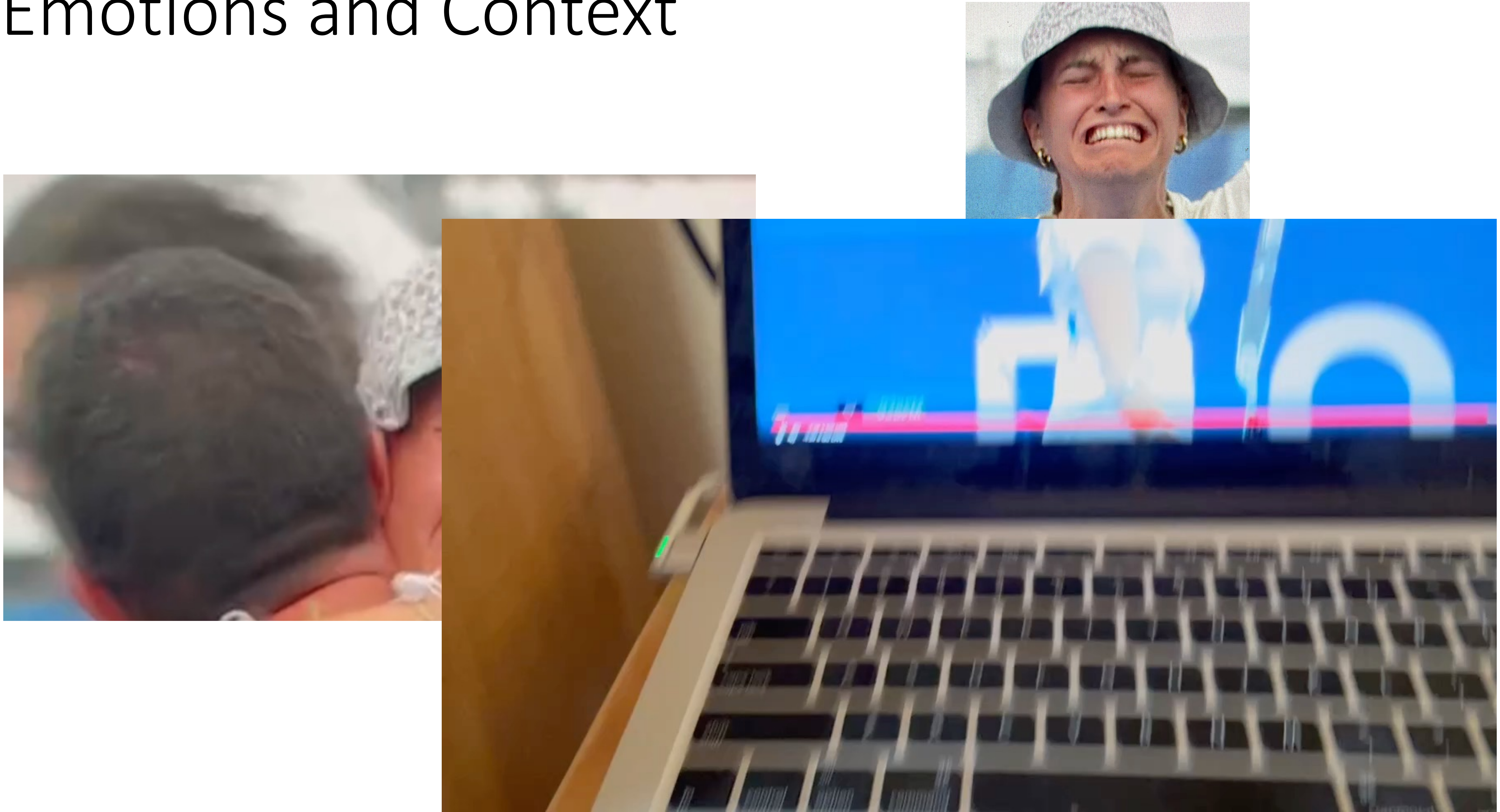
Now:

- IEEE Transactions on Affective Computing
- International Conference on Affective Computing and Intelligent Interaction (ACII)
  → https://acii-conf.net/
- Increasing awareness that emotions play a significant role in human intelligence
- but, still don’t have “emotional machines” - why not?

Facial “Expressions of Emotion” and the Ekman Program

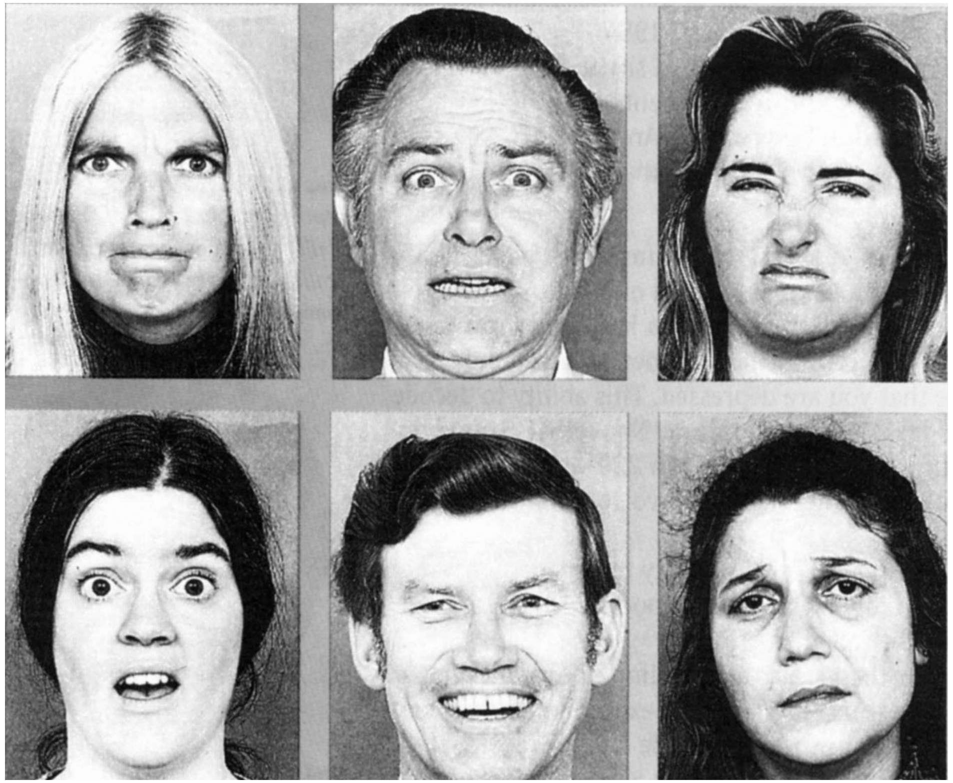


Emotions and Context

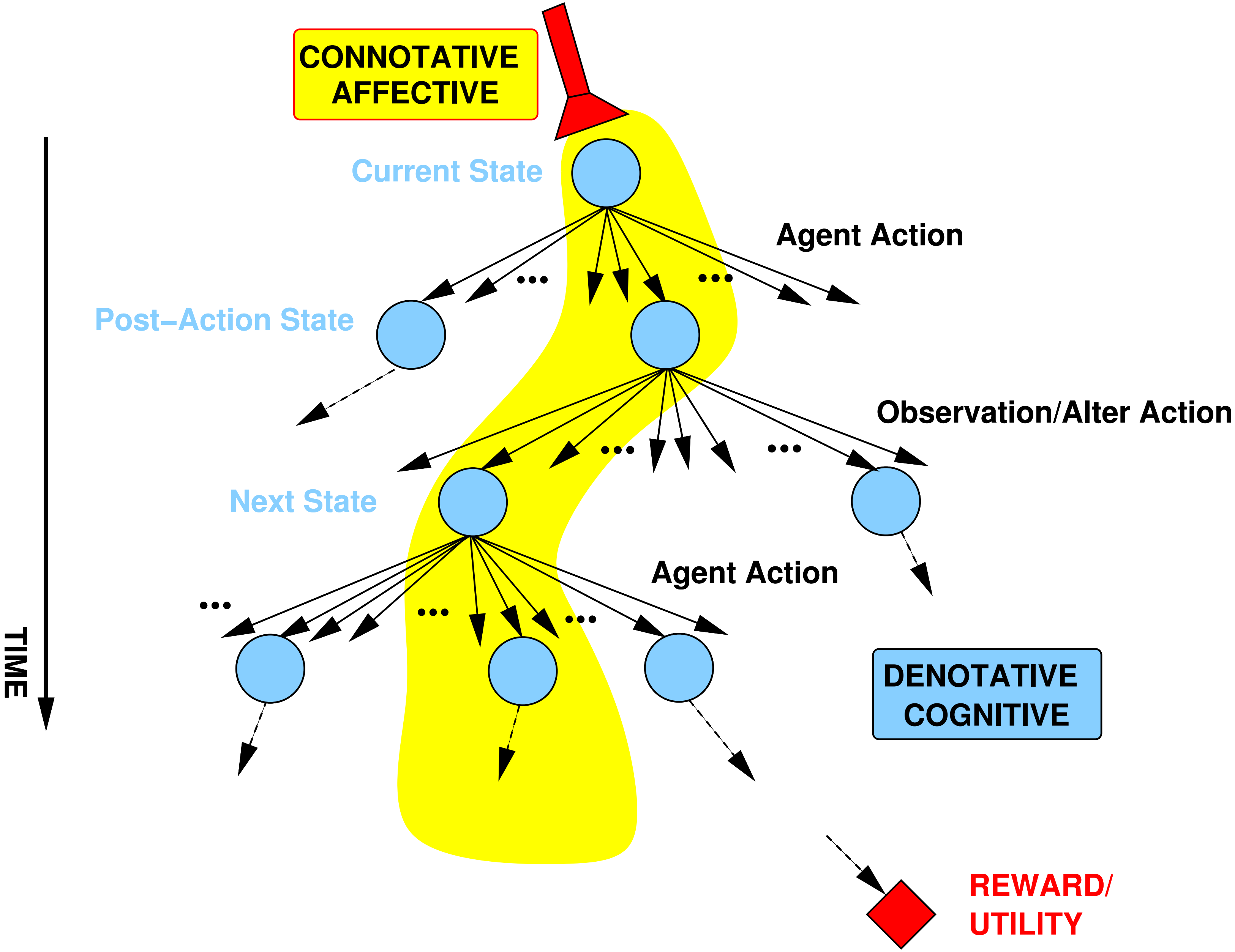


context sensitive,
emotionally unaware

context insensitive,
emotionally aware



Human-Centered A.I. using a Socio-Emotional "flashlight"



CONNOTATIVE AFFECTIVE

Current State

Agent Action

Post–Action State

Observation/Alter Action

Next State

Agent Action

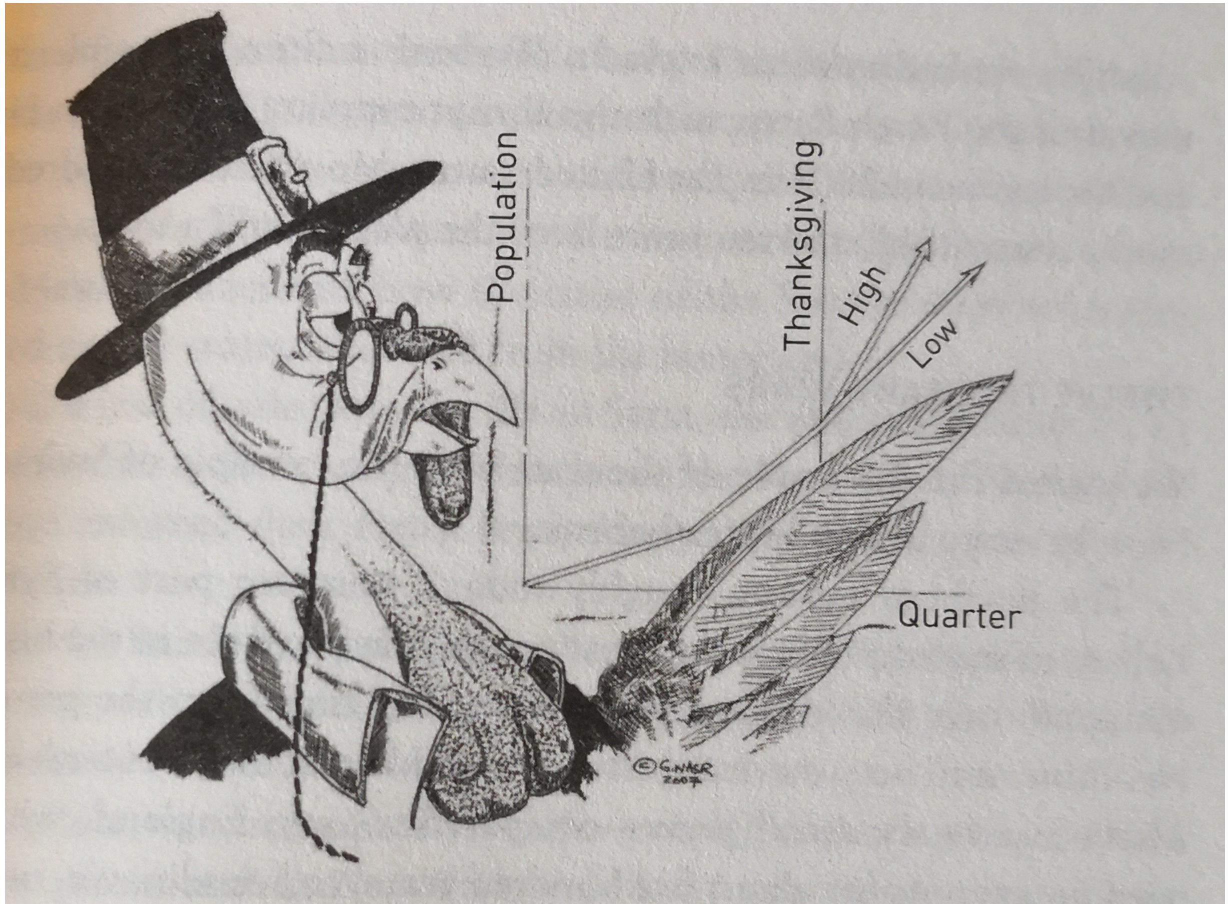
TIME

DENOTATIVE COGNITIVE

REWARD/ UTILITY

White Swans …

Black Swan Events



From N.N. Taleb “AntiFragile” p. 93

Monoculture…

Prussia 1765-1800: took 100 years before the damage was apparent

(J.C.Scott *Seeing like a State*, 1998)

Efficient Crickets…

Computers and sensors will track everything from environmental factors like heat and humidity, to cricket growth, waste production, and even check for sick or dead insects.

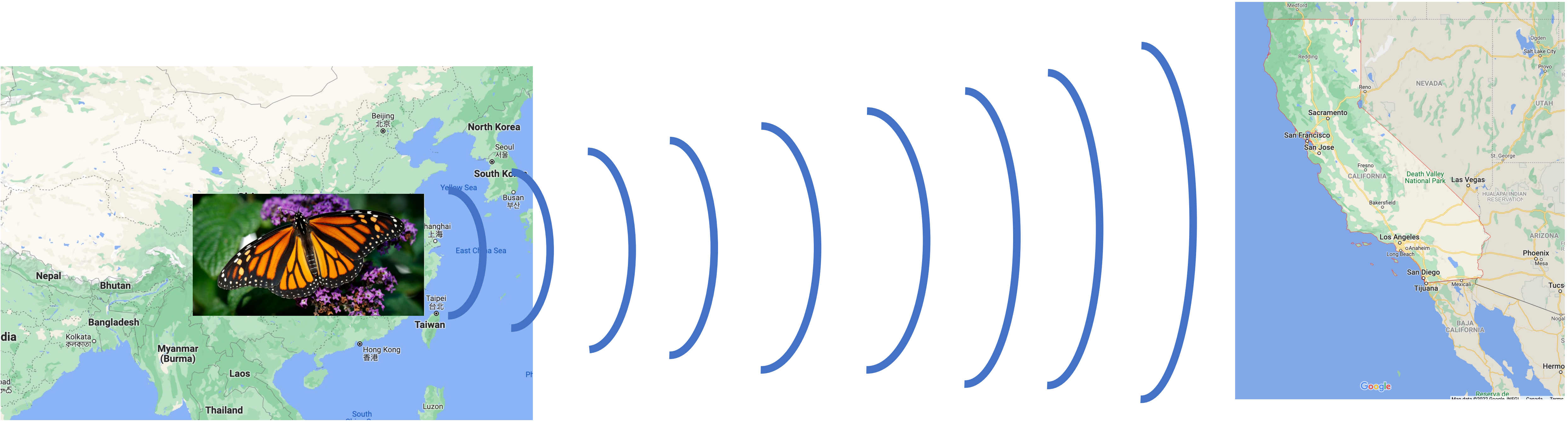
"You need a lot of data with AI, and with living organisms how do I get enough examples of sick crickets for the algorithm to learn? You obviously don't want to create that for ethical reasons ... so it is a technical hurdle we'll have to overcome."

Fernandez, whose company has already partnered with German automotive giant Audi and U.S. defence contractor Lockheed Martin, acknowledged that tracking the health and well-being of living, breathing crickets will be different from analyzing automotive or aerospace components.

"Using artificial intelligence to raise better crickets and help save the world"
Waterloo Record By [James Jackson]() Record Reporter Mon., Feb. 22, 2021

Complexity: Black Swans arise

Many interacting parts, long distance, path-based effects



Beam Search

Need a small enough beam to see what you are looking at …

… but a big enough beam not to miss the black swan.

Bayesian Reinforcement learning

- Normal Reinforcement Learning (RL):
  - Trying to learn model (function) $action = m(history)$ from a class of models
  - Model based: $[action, P(s')] = \sum_S f(history, s)$ where *s*=latent states
  - Use a fixed class of *m*, and latent space,

    E.g. say deep neural networks with some architecture
- Bayesian RL:
  - Learn the distribution over m
  - $[action, P(s', m')] = apply(m, history)$
  - Bayesian model comparison
  - Handles non-stationary environments
  - Different variants:
    - meta-learning, learning to learn, transfer learning, active/sophisticated inference
  - Now you (and everybody) can change flashlights!

Bayesian Learning

To learn the distribution over models, m,

we compute:

d: data

M: class of models

$$P(m|d) = \frac{P(d|m)P(m)}{\sum_m P(d|m)P(m)} \quad m \in M$$

P(d): Evidence, partition function or $e^{-\text{free energy}}$ is hard to compute

sum over all models, m ∈ M

© Steve Winter/National Ge

free energy

realization ↔ expectation

Human Expectations vs Realizations

Out of distribution learning

- (Social) Emotionally based
- Handles surprises that arise from complexity
- Cooperative behavior is emphasized
- Emotional alignment
- Critical for human-centered AI
- Example domains: Assistive Technology, Climate Change, Recommender Systems, Robots, General Game Playing, Politics,…
- Implementation: BayesACT

Artificial Intelligence: where next?

- intelligence ≠ rationality
- it’s more about how to be a **flexible, free & innovative individual** in a **equal, robust and secure society.**
- “heuristic” social intelligence
- emotions encode a social order (the “flashlight”) that allows us to work in a society

~~Can artif [illegible] intelligent as an indivi [illegible]~~

Can artificial agents become members of a human society?

Human Society: a complex system

- Build AI to work with people in a way that enhances collaboration and cooperation and preserves freedom/independence?

- Is emotion the social glue that holds groups together, yet allows innovation?

- Can AI use emotions to promote human benefits (e.g. UN Sustainable Development Goals)?

"…you have to build a relationship of trust first, then you can provide assistance." – Pete [illegible], [illegible]uen Mobile Health, Kitchener

Recent and Current Collaborators:

Alex Mihailidis (Toronto)
Tobias Schroder (Potsdam)
Julie M. Robillard (UBC)
Chris Nugent (Ulster)
Luke Chen (Ulster)
Heather Love (Waterloo)
Shehroz Khan (Toronto)
Charlene Chu (Toronto)
Kathryn Lively (Dartmouth)
Mary Step (Cleveland State)
Enrique Sucar (INAOE)
Sarel Van Vuuren (Colorado)
James Tung (Waterloo)
Rosalie Wang (Toronto)
Neil MacKinnon (Guelph)
Alexandra Konig (INRIA)
Stefan Teipel (Rostock)
Francois Michaud (Sherbrooke)
Linda Francis (Cleveland State)
Kimberly B. Rogers (Dartmouth)

Current Students and Postdocs (Waterloo):

Renee Leung
Jess Gano
Blake Vanberlo
Christopher Risi
Heather Eustace
Alyssa Clarkson
Joshua Jung
Aarti Malhotra
Freya Zhang
Gabby Chan
Wasif Khan

Funding

- -Trans-Atlantic Partnership (TAP)
- -Cheriton, Graham Trust (Waterloo)
- -Natural Science and Engineering Research Council, (NSERC, Canada)
- -Social Science and Humanities Research Council, (SSHRC, Canada)

Coordinates

bayesact.ca

jhoey@uwaterloo.ca

@drjessehoey