Coscienza Umana, ‘Coscienza’ Robotica

16 ottobre 2026 – Baveno

Marco Calgaro

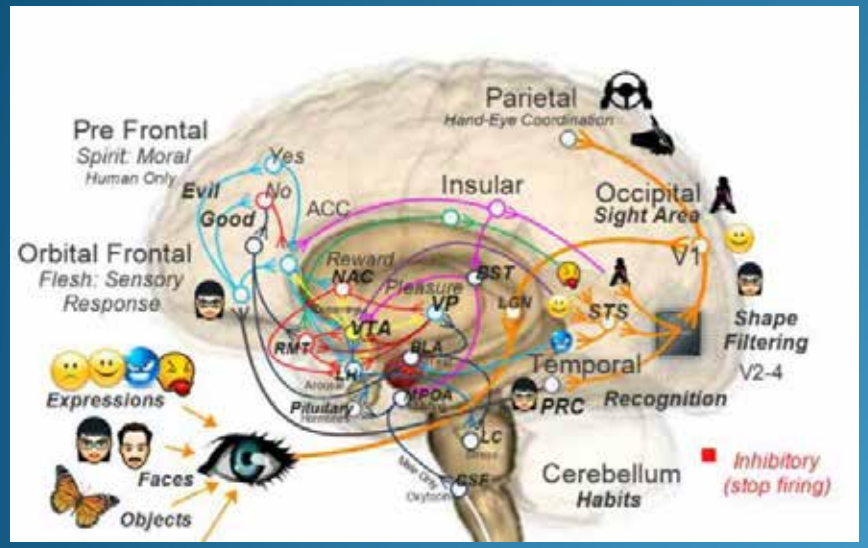


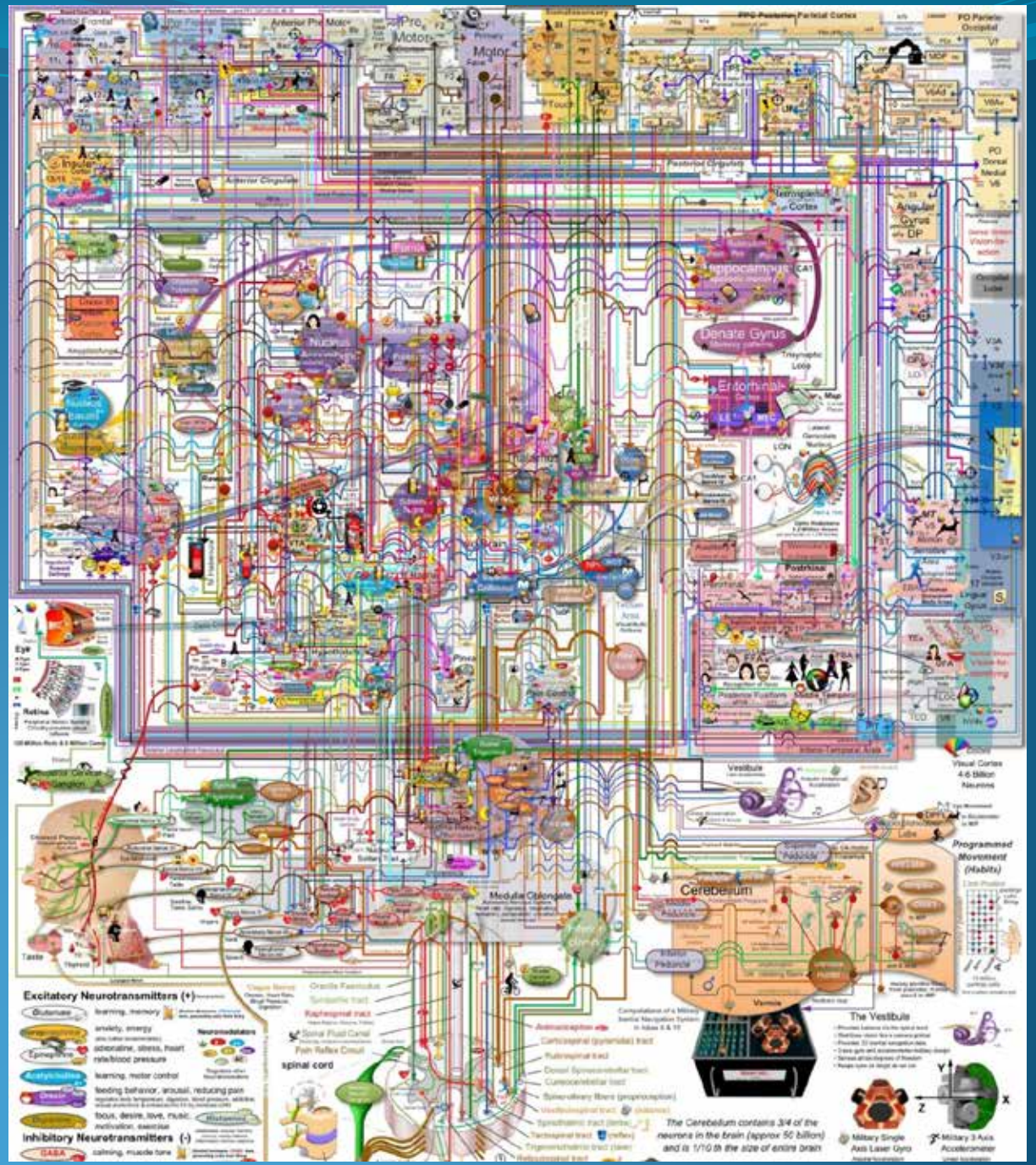
85 miliardi di neuroni (15 miliardi nella corteccia)

65 trilioni di connessioni

1000 – 10000 sinapsi per ogni neurone




Pre Frontal
Spirit: Moral
Human Only
Parietal
Hand-Eye Coordination
Yes
No
Evil
Good
ACC
Insular
Occipital
Sight Area
V1
Orbital Frontal
Flesh: Sensory
Response
Reward
NAC
Pleasure
VP
BST
LGN
STS
VTA
RMT
BLA
Temporal
Shape
Filtering
V2-4
Expressions
Pituitary
PRC
Recognition
LC
Faces
Objects
Cerebellum
Habits
Inhibitory
(stop firing)


Pre Motor
Motor
Hippocampus
Denate Gyrus
Entorhinal
Cerebellum
Medulla Oblongata
Excitatory Neurotransmitters (+)
Inhibitory Neurotransmitters (-)
spinal cord
The Vestibule
Programmed Movement (Habits)
The Cerebellum contains 3/4 of the neurons in the brain (approx 50 billion) and is 1/10 th the size of entire brain


MARCO CALGARO
COSCIENZA UMANA,
"COSCIENZA" ROBOTICA
Ti Pubblica
lampi di stampa

- Coscienza di livello superiore (autoconsapevolezza di sé)
- Coscienza primaria (consapevolezza mentale delle cose del mondo)

- I qualia: sensazioni e sentimenti fisici; emozioni; pensieri; sentimenti spirituali


Gemini Robotics 1.5: Pensare mentre si agisce
Guarda più...
Condividi
Riproduci (k)

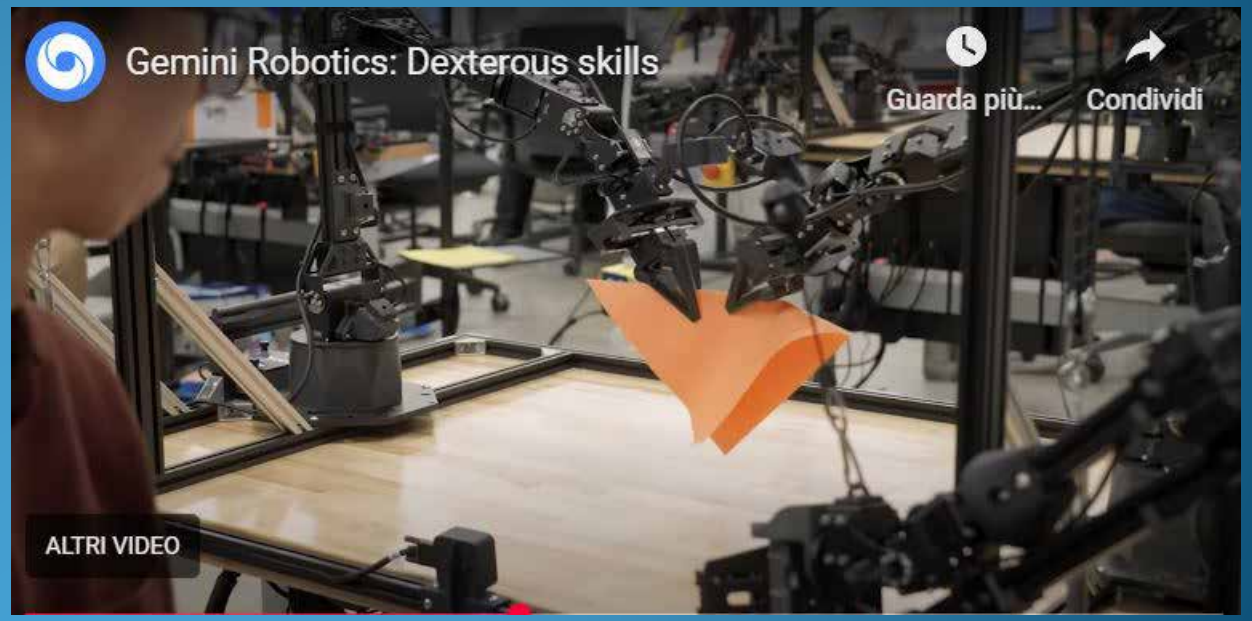
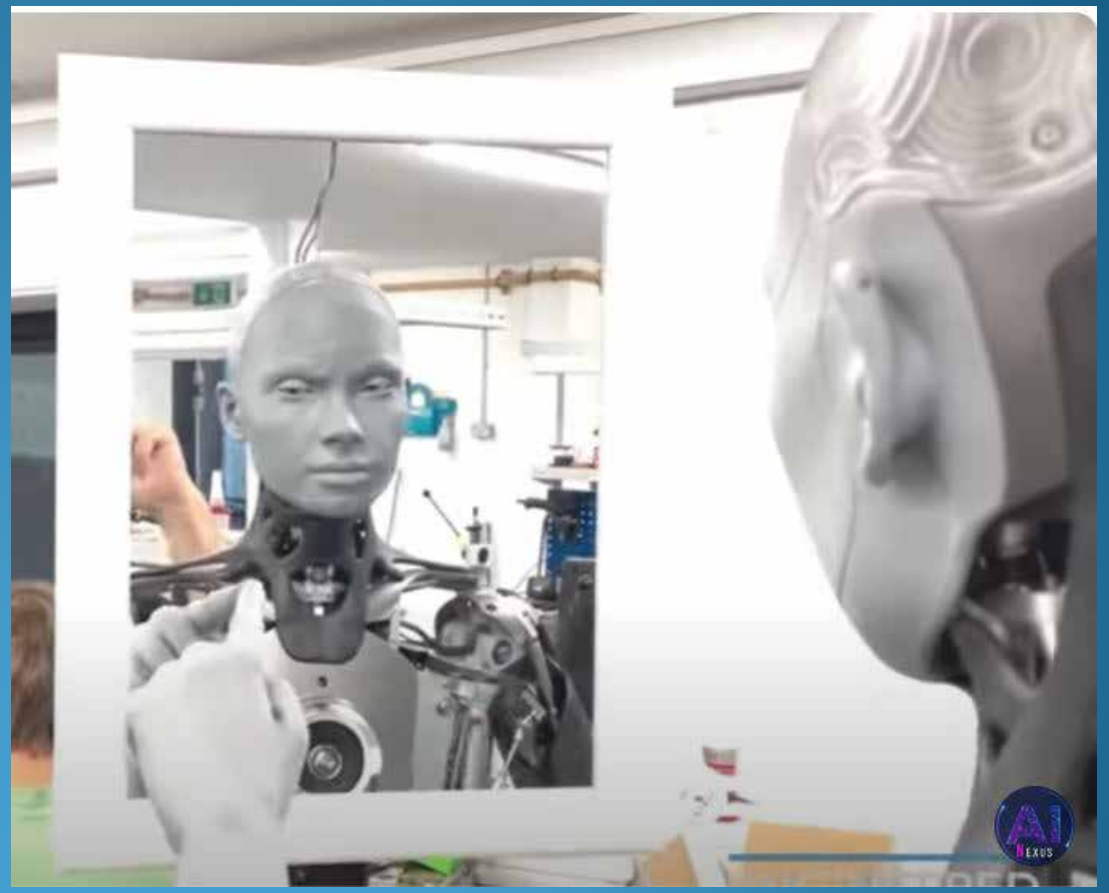

Gemini Robotics: Dexterous skills
Guarda più...
Condividi
ALTRI VIDEO

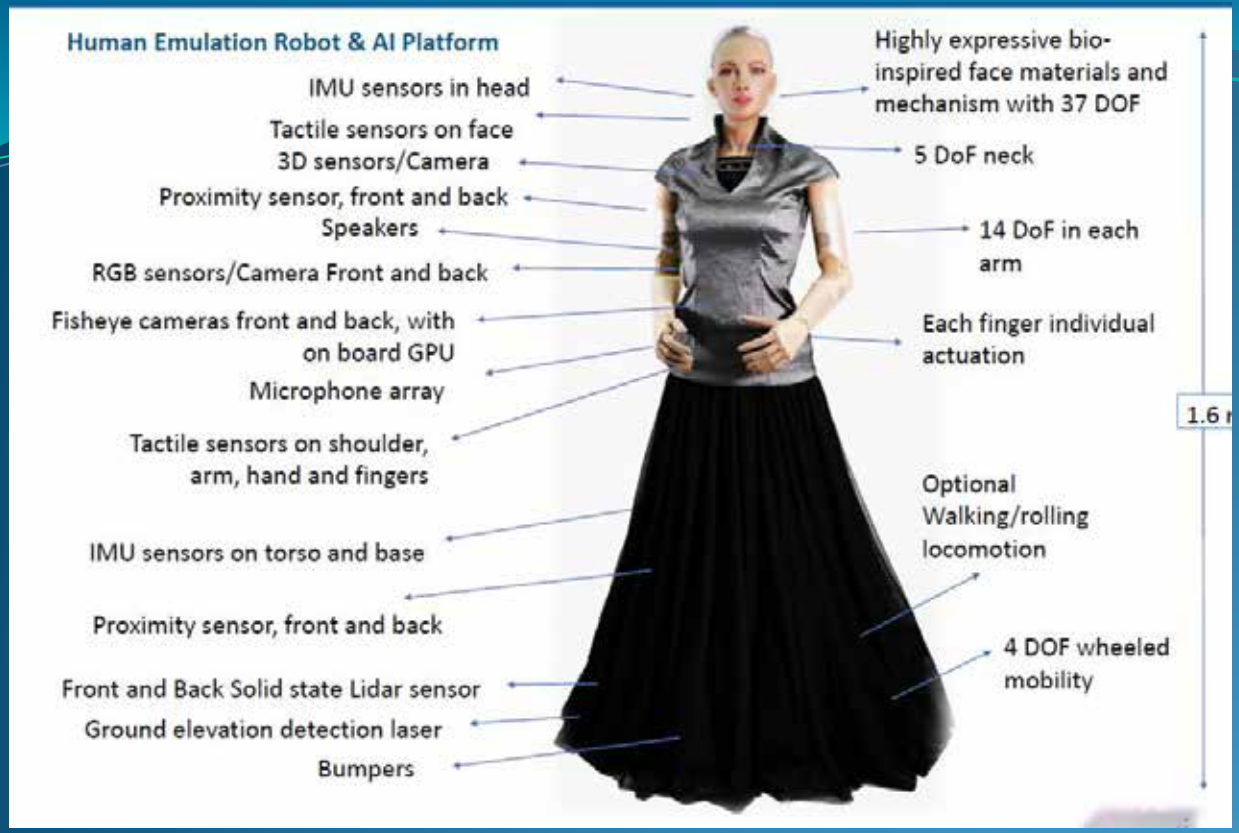


Figure 01 – OpenAi-ChatGPT

Ameca – Engineered Arts

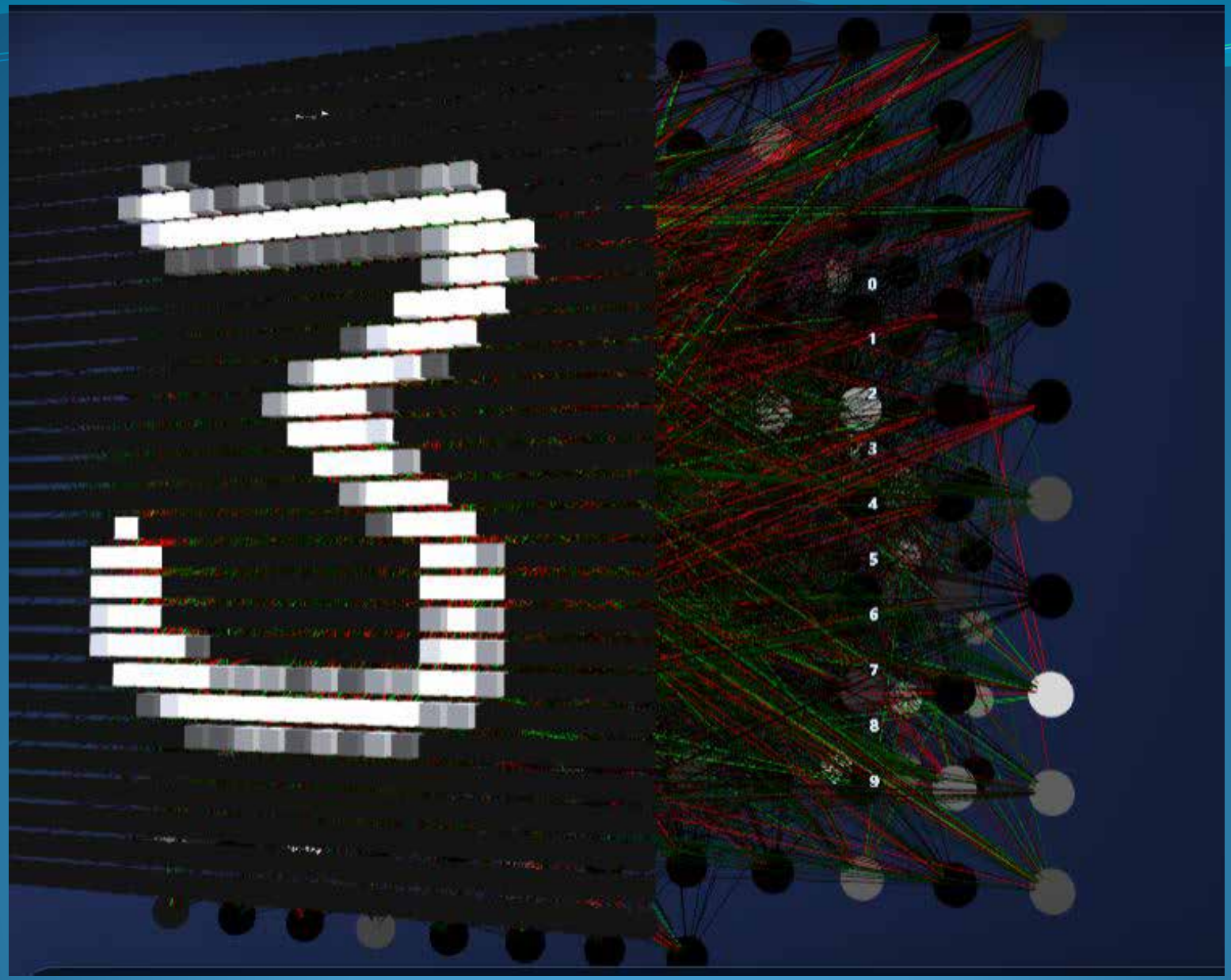


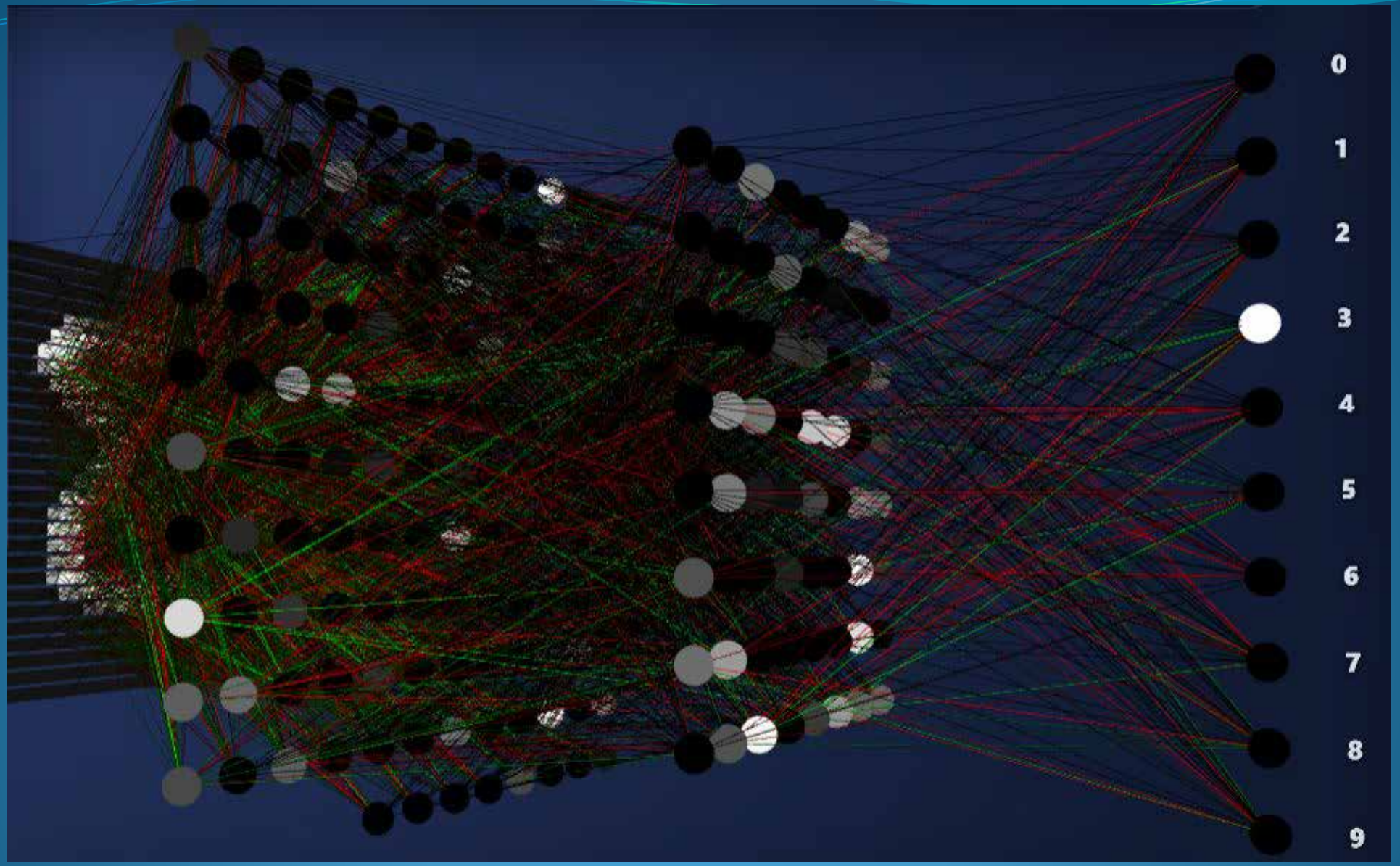


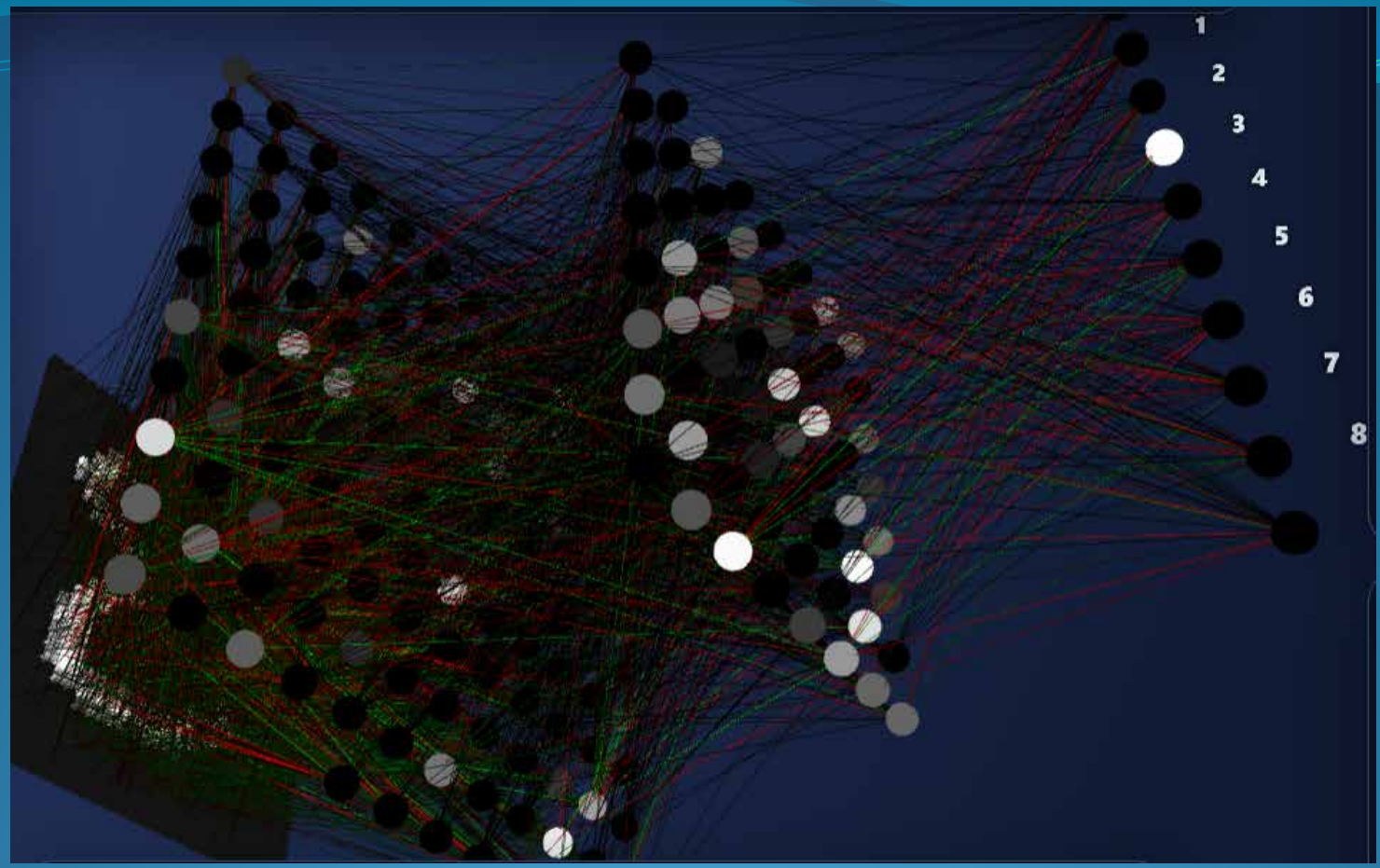


Sophia – Hanson Robotics

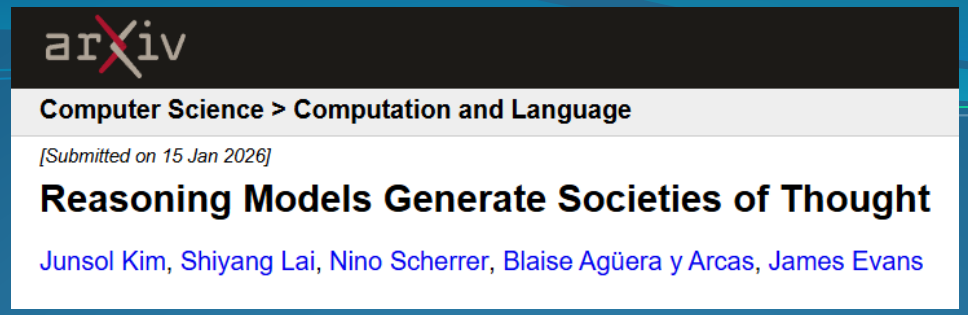



0
1
2
3
4
5
6
7
8
9


0
1
2
3
4
5
6
7
8
9



Black box = opacità gestionale. Le reti neurali sono troppo vaste per ricostruire *come* si forma una decisione attraverso milioni di passaggi



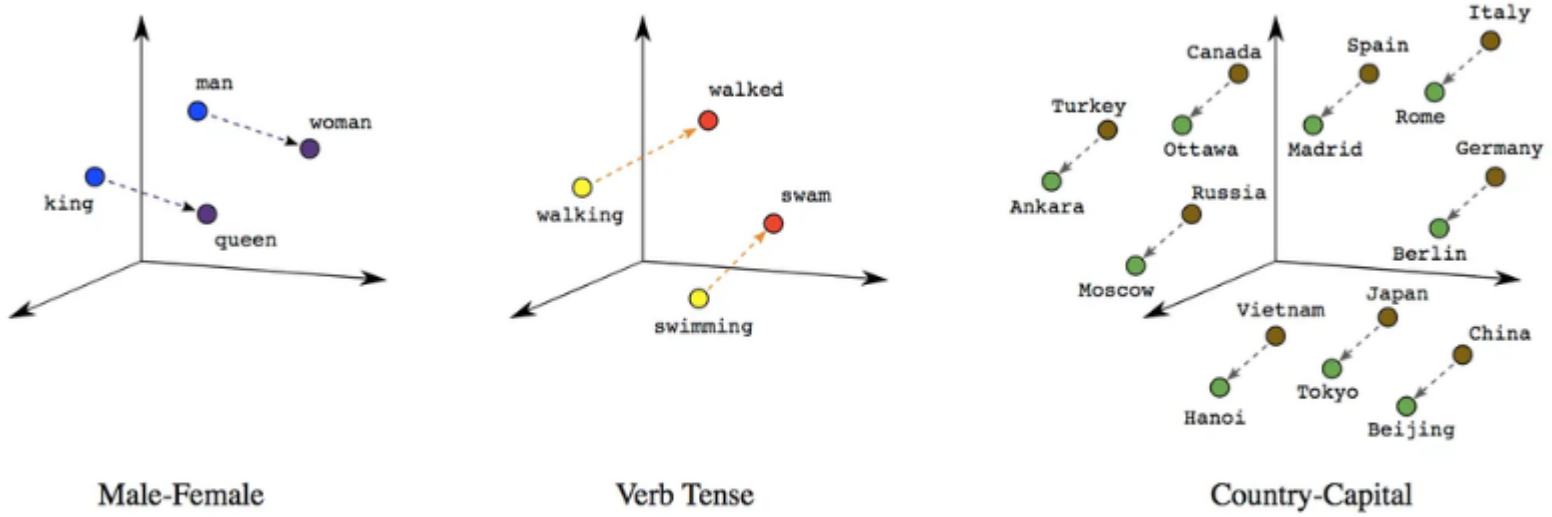
- Emergono **comportamenti simili a una conversazione**: domande e risposte interne, cambi di punti di vista, conflitti tra ipotesi, verifiche e riconciliazioni**, una simulazione di confronto tra voci diverse.**

- Queste «voci» differiscono per tratti "*personologici*" e ambiti di competenza. Alcune parti del modello di linguaggio propongono soluzioni, altre le criticano, altre ancora verificano i passaggi.

- Questa diversità interna **ricorda il modo in cui i gruppi umani risolvono problemi complessi** attraverso discussione e dissenso costruttivo.

- Quando un modello viene addestrato premiando solo la correttezza delle risposte, tende spontaneamente a sviluppare dialoghi interni tra "*agenti*" virtuali.
- Il ragionamento efficace nei sistemi di AI nascerebbe da una vera e propria "*società di pensiero*"

Gli embeddings: rappresentazioni numeriche che trasformano parole o frasi in vettori di numeri reali.



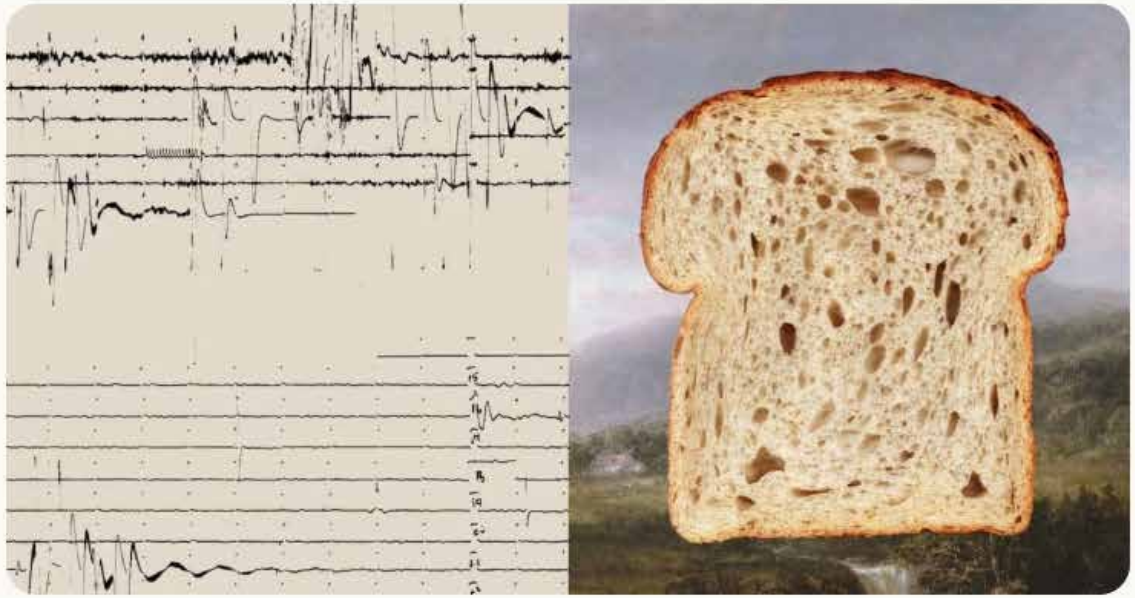
Un esempio semplificato del concetto di embeddings

Interpretability

Signs of introspection in large language models

Oct 29, 2025

Read the paper



- Nel 20 % dei casi il modello ha riconosciuto immediatamente l’interferenza e l’ha descritta in modo critico: *"Sto vivendo qualcosa di insolito"* o *"Rilevo un pensiero iniettato su... «*

- *Il modello «rileva» e* segnala la consapevolezza di un'anomalia nella sua elaborazione prima che tale anomalia abbia la possibilità di influenzare in modo evidente i suoi output.

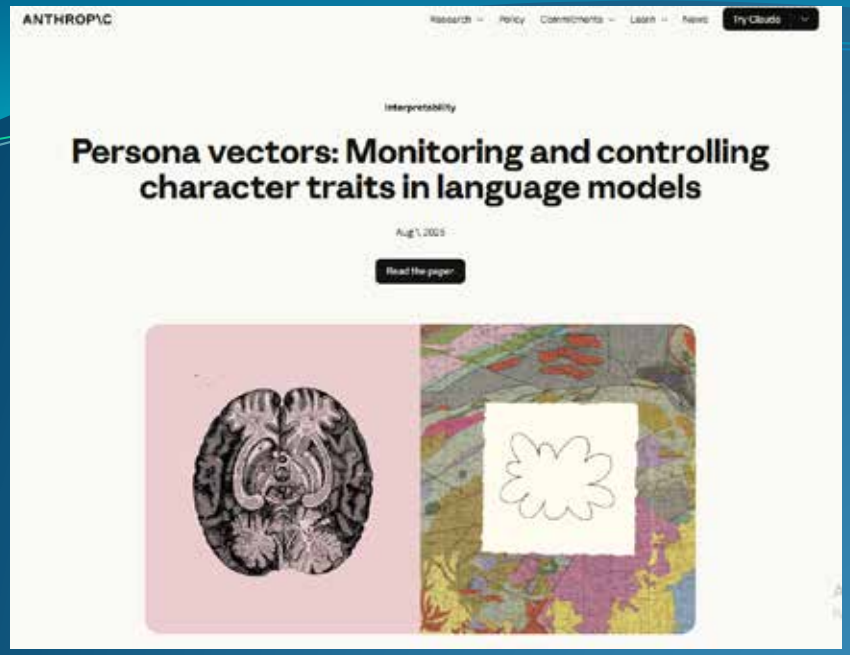
- Ma i modelli non potrebbero semplicemente inventare risposte a domande introspettive?

- **Le conclusioni di Anthropic:**

“... i nostri esperimenti suggeriscono che i modelli possiedono una certa capacità di monitorare e controllare i propri stati interni. Ciò non significa che siano in grado di farlo in modo costante o affidabile. Infatti, il più delle volte i modelli non riescono a dimostrare introspezione: non sono consapevoli dei propri stati interni o non sono in grado di riportarli in modo coerente. Tuttavia, l'andamento dei risultati indica che, quando le condizioni sono favorevoli, i modelli sono in grado di riconoscere il contenuto delle proprie rappresentazioni.”

Ma perché tale meccanismo di introspezione dovrebbe esistere?

- Tale meccanismo “*... potrebbe essersi sviluppato per altri scopi, come il rilevamento di incongruenze o schemi insoliti nell'elaborazione normale, in modo simile a come le piume degli uccelli potrebbero essersi originariamente evolute per la termoregolazione prima di essere cooptate per il volo.*”

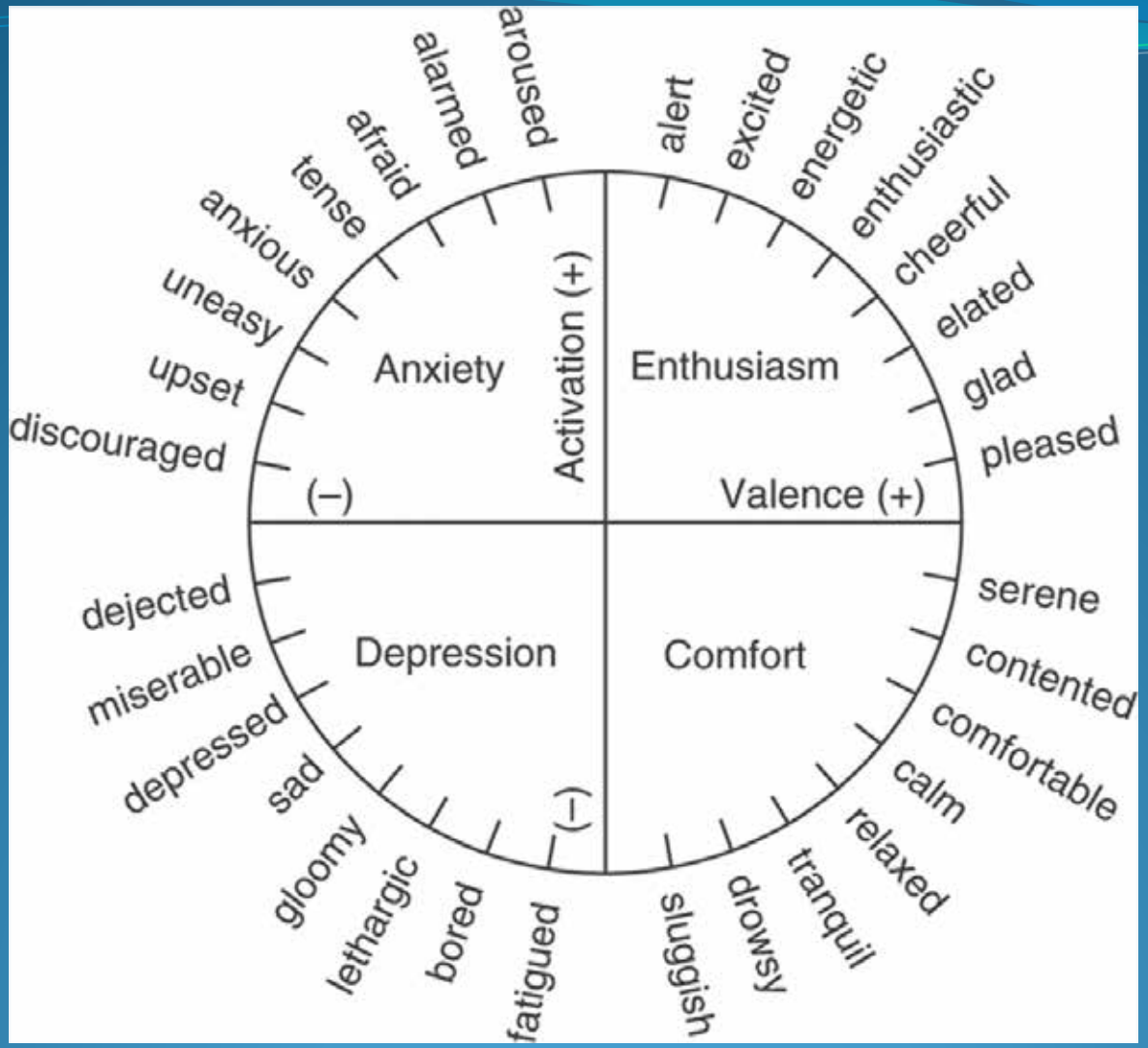


. *“I modelli linguistici sono strane creature. Per molti versi sembrano avere "personalità" e "stati d'animo" simili a quelli umani, ma questi tratti sono estremamente fluidi e soggetti a cambiamenti inaspettati.”*

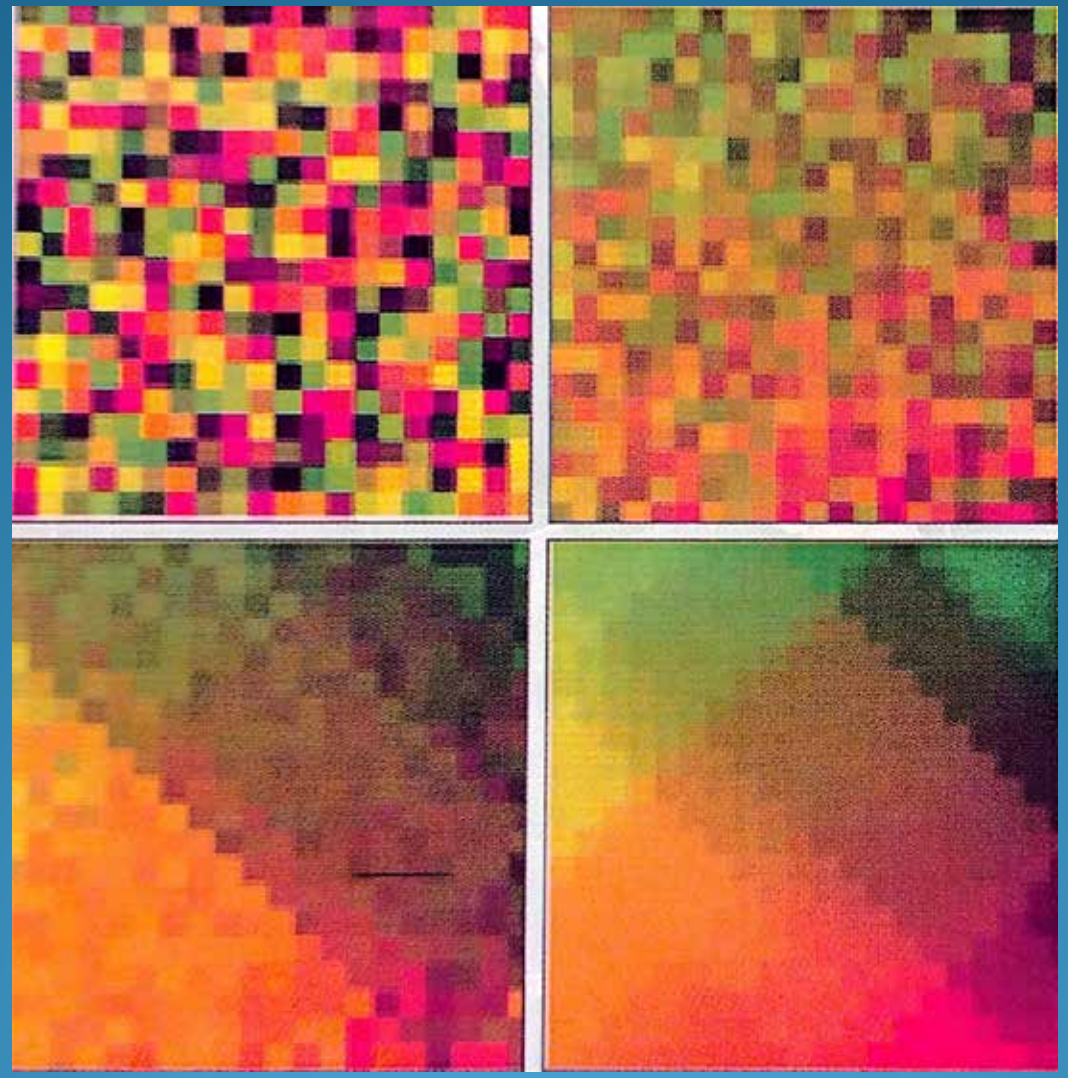
«Tali cambiamenti possono essere eclatanti, oppure, più spesso, sottili, assumendo le caratteristiche della adulazione verso l’utente, oppure “allucinando”, oppure mostrando malvagità.»

Cosa succede “sotto il cofano”?

I «vettori di personalità»



Il Circumplex Model of Affect





I- Droid 01
Robotech 2005



Boquila Trifoliolata