

# Gestione della Conoscenza per i Servizi Digitali



Antonio Lieto

Università di Salerno, IT  
ICAR-CNR, Palermo, IT

# Temi del corso

Introduzione ai paradigmi di modellazione dell'**Intelligenza Artificiale** (neurale-simbolica-ibrida)

Paradigma Simbolico: **Logica, Inferenza e Sistemi di Rappresentazione e Gestione della Conoscenza**

**Ontologie e Tecnologie Semantiche** (con parti di laboratorio su costruzione, immagazzinamento e interrogazione di modelli)

# Esame

Progetto di gruppo (min 2 max 4 persone) sulla parte laboratoriale da consegnare e discutere

Orale individuale su tutto il programma

## Testi e Materiale

In Inglese

Materiale diffuso a lezione (slides, manuali per software). Disponibili su <https://www.antonioletto.net> (andare in Didattica)

Libro di testo da riferimento per tutta la parte teorica

## Testi e Materiale

In Inglese

Materiale diffuso a lezione (slides, manuali per software)

Libro di testo da riferimento per tutta la parte teorica

**Lieto A, 2021, Cognitive Design for Artificial Minds, Routledge/  
Taylor & Francis, London/New York.**

**<https://www.routledge.com/Cognitive-Design-for-Artificial-Minds/Lieto/p/book/9781138207950>**

# Cognitive Design for Artificial Minds

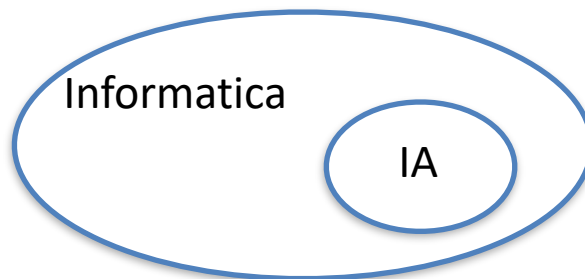
Antonio Lieto



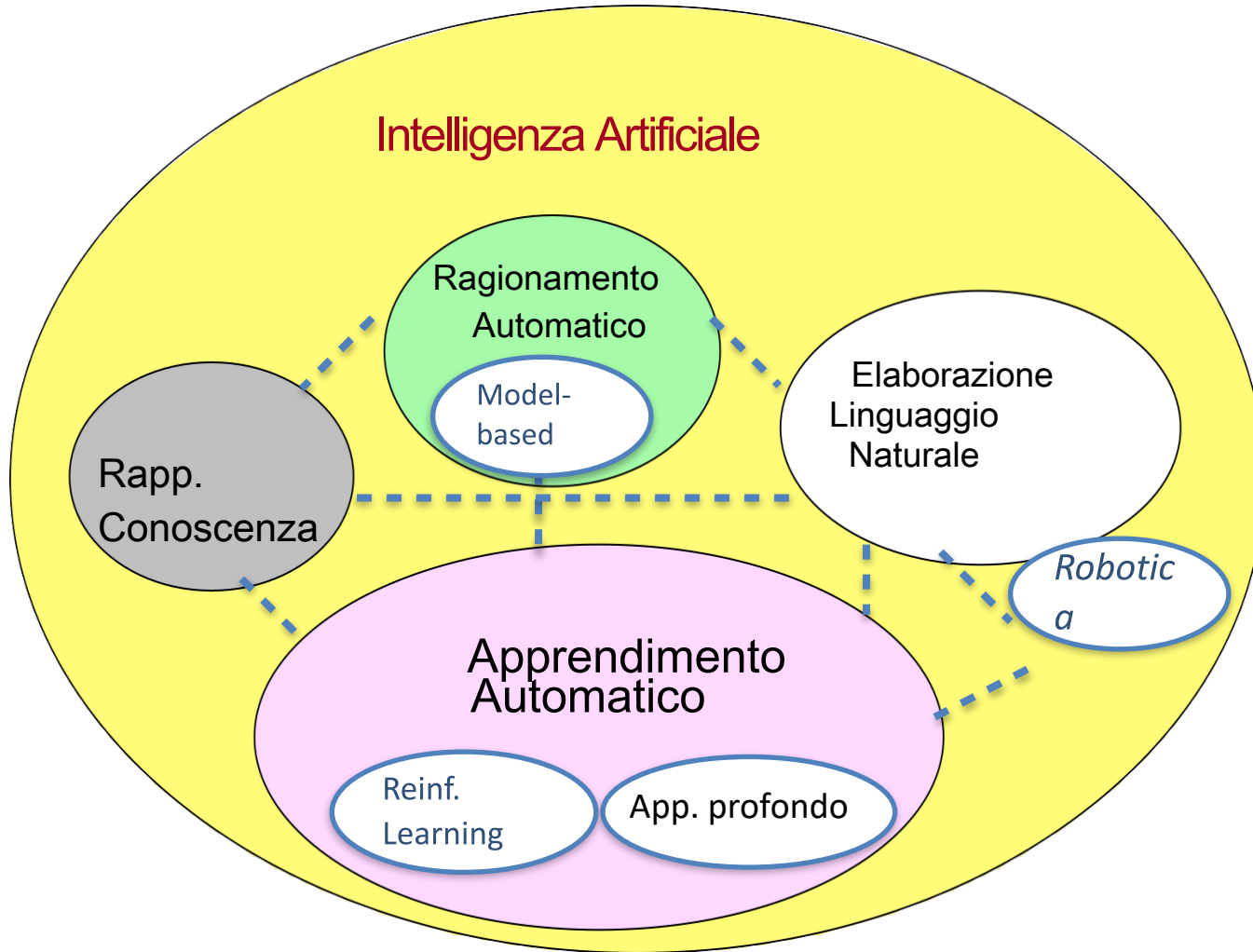
# Una definizione

«L'**intelligenza artificiale** è una branca dell' **informatica** che studia i fondamenti teorici, le metodologie e le tecniche che consentono la progettazione di sistemi hardware e software capaci di fornire a un calcolatore prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana.»

(Somalvico, 2003)



## Tante Aree di ricerca in Intelligenza Artificiale (elenco parziale)





# Intelligenza Artificiale

L'IA è contemporaneamente scienza e tecnica, disciplina di frontiera in cui si incontrano diversi domini del sapere: logica, informatica, psicologia, neuroscienze, filosofia...

Attiene a tutti i campi dell'informazione e trova applicazione in tutti i settori

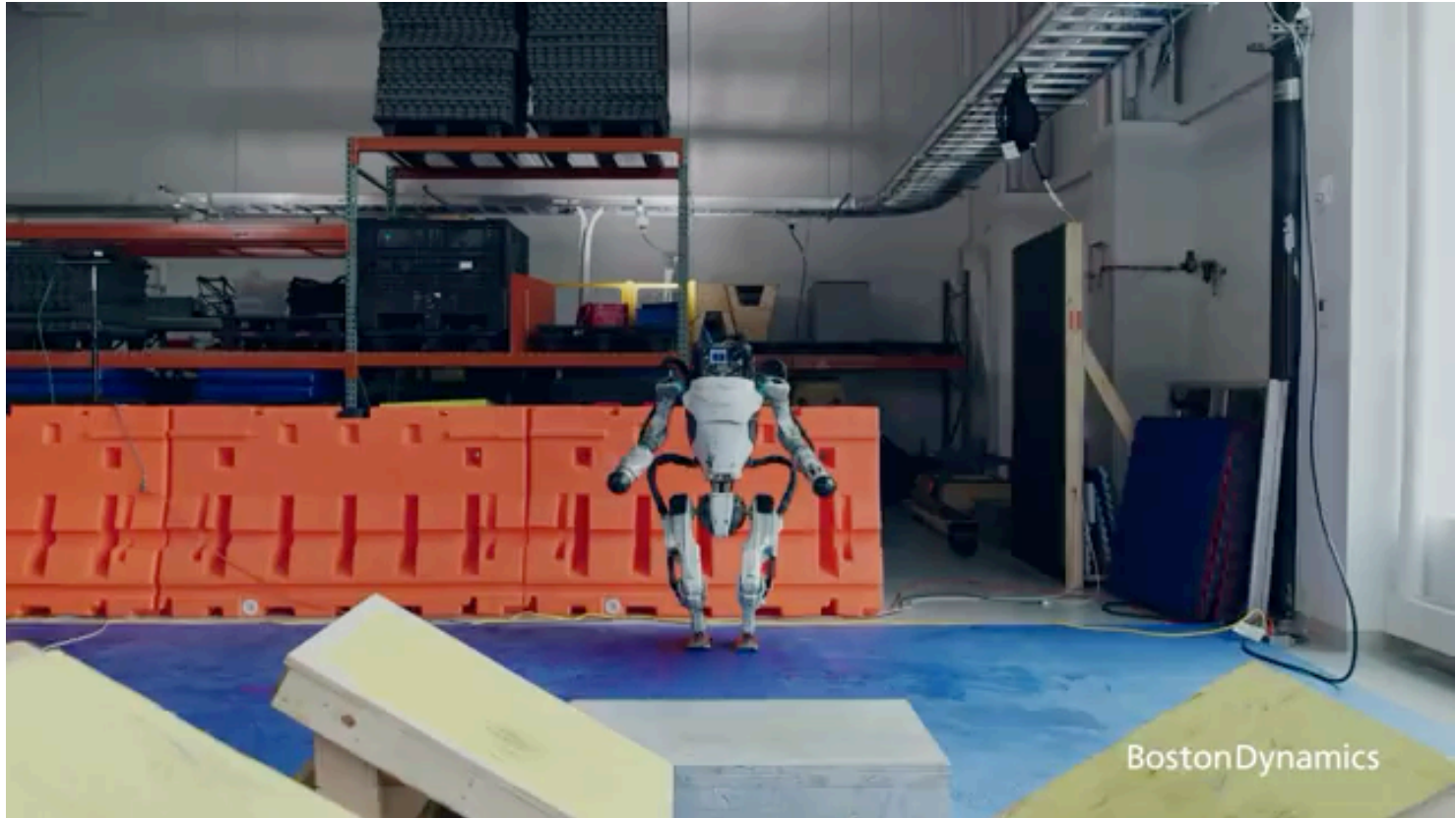
# Es. Veicoli a guida autonoma

## Example: Car driving



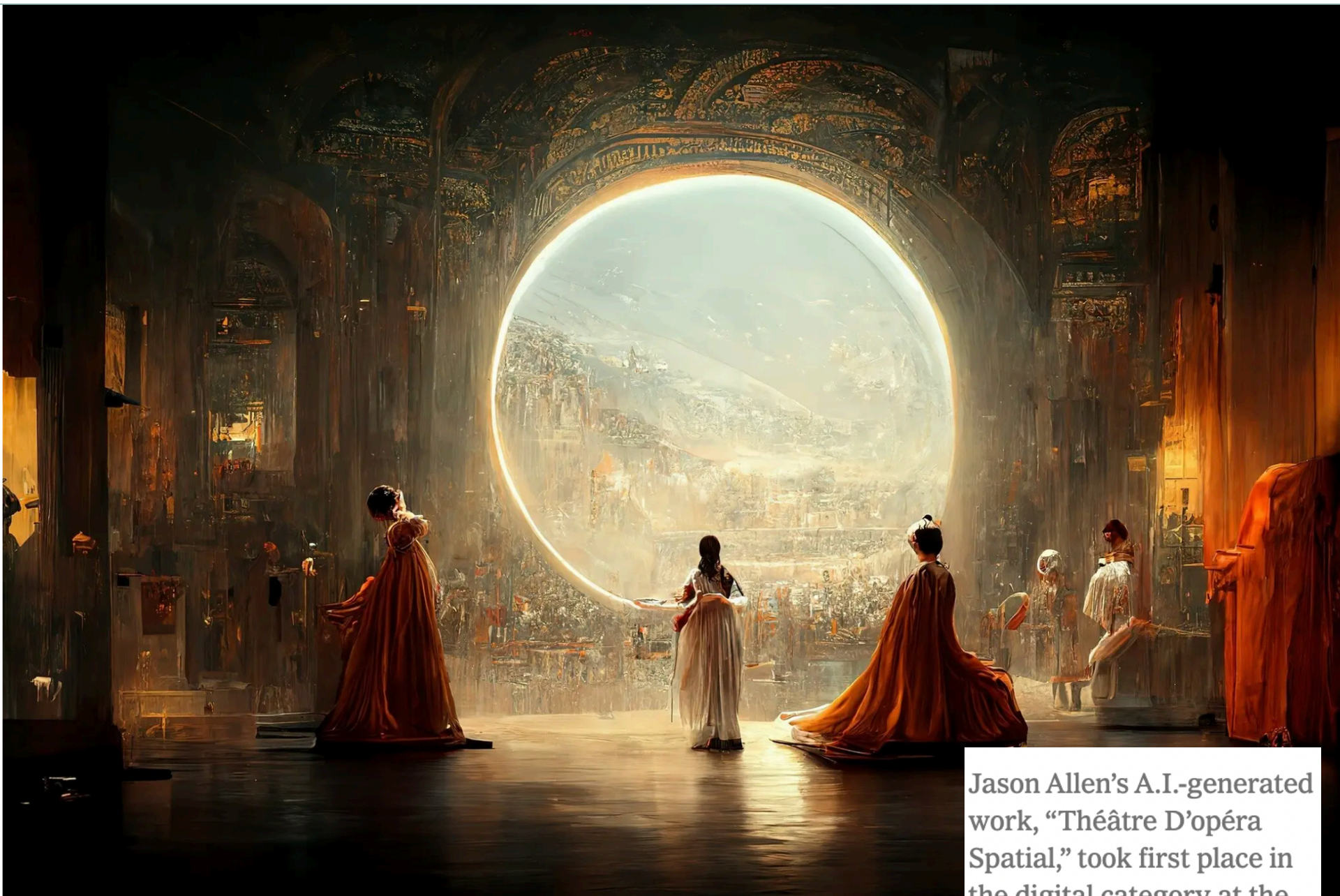
Actually a “Perception to Action” system. The visual perception and control system is a Deep learning architecture trained end to end to transform pixels from the cameras into steering angles. And this car uses regular color cameras, not LIDARS like the Google cars. It is watching the driver and learns.

<https://www.youtube.com/watch?v=YuyT2SDcYrU>



**The  
Guardian**





Jason Allen's A.I.-generated work, "Théâtre D'opéra Spatial," took first place in the digital category at the Colorado State Fair.

SYSTEM PROMPT  
(HUMAN-WRITTEN)

*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

MODEL COMPLETION  
(MACHINE-WRITTEN,  
10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

# GPT-3

# Intelligenza Artificiale

Dal punto di vista storico l'IA si manifesta con un duplice profilo:

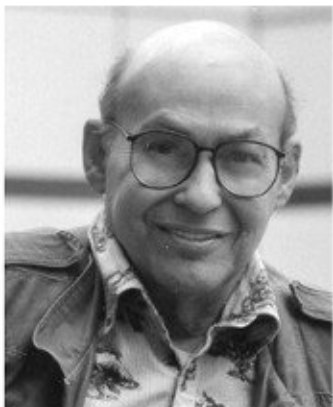
- *disciplina ingegneristica* (costruire macchine capaci di svolgere compiti intelligenti integrando componenti diverse)
- *disciplina psicologica* (costruire macchine capaci di esprimere le caratteristiche essenziali dell'attività cognitiva umana che siano in grado di spiegare meccanismi complessi che riguardano il funzionamento di alcuni fenomeni della nostra "mente": es. ragionamento, apprendimento, percezione etc.)



# 1956 Dartmouth Conference: The Founding Fathers of AI



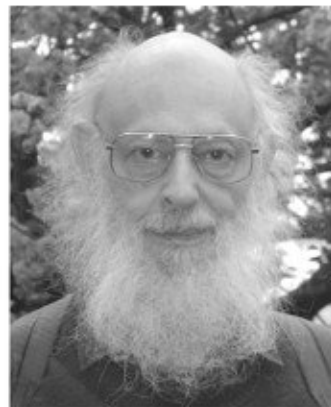
**John MacCarthy**



**Marvin Minsky**



**Claude Shannon**



**Ray Solomonoff**



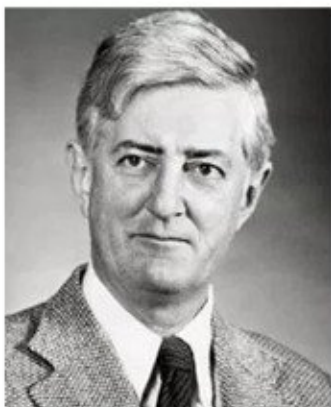
**Alan Newell**



**Herbert Simon**



**Arthur Samuel**



**Oliver Selfridge**



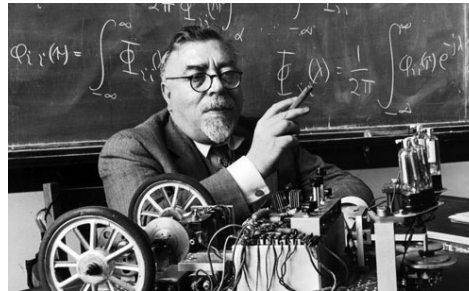
**Nathaniel Rochester**



**Trenchard More**

# IA e cibernetica

Norbert Wiener **“Cybernetics: Or Control and Communication in the Animal and the Machine”**.



Idea: costruire modelli meccanici in grado di simulare il comportamento adattivo di sistemi naturali

(Cordeschi, 2002): **“L’intuizione principale della cibernetica è stato quello di proporre una prospettiva unificata allo studio di organismi biologici e macchine” (all’epoca intese soprattutto come sistemi meccanici).**

# Trend storici

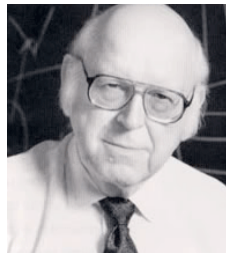
## AI dei primordi

Simbolica  
Ispirazione Cognitiva

## IA oggi

focus su task specifici  
ibrida o connessionista

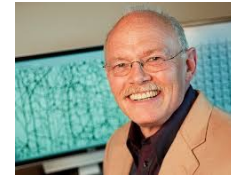
A. Newell



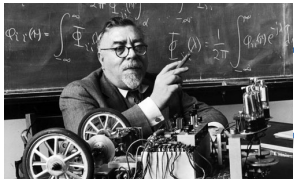
M. Minsky



J. McClelland



mid'80s



N. Wiener



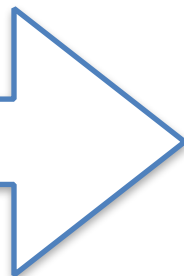
H. Simon



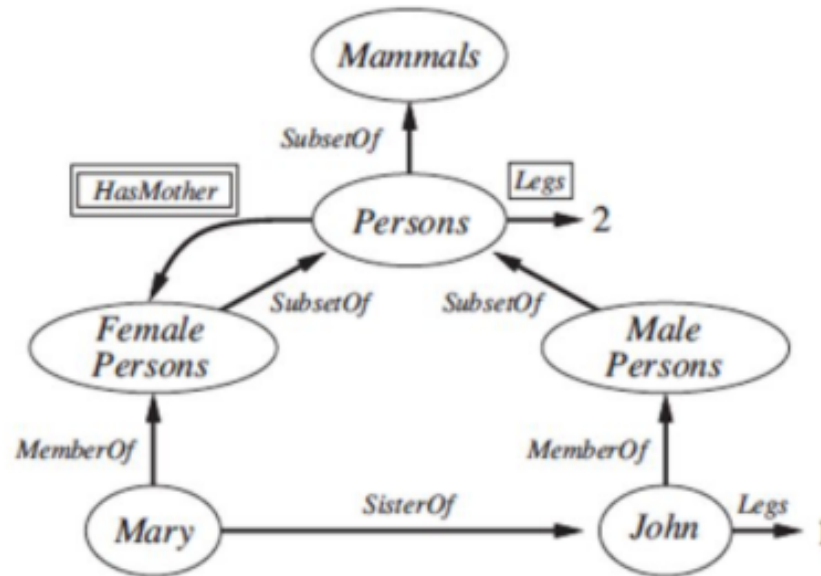
R. Shank



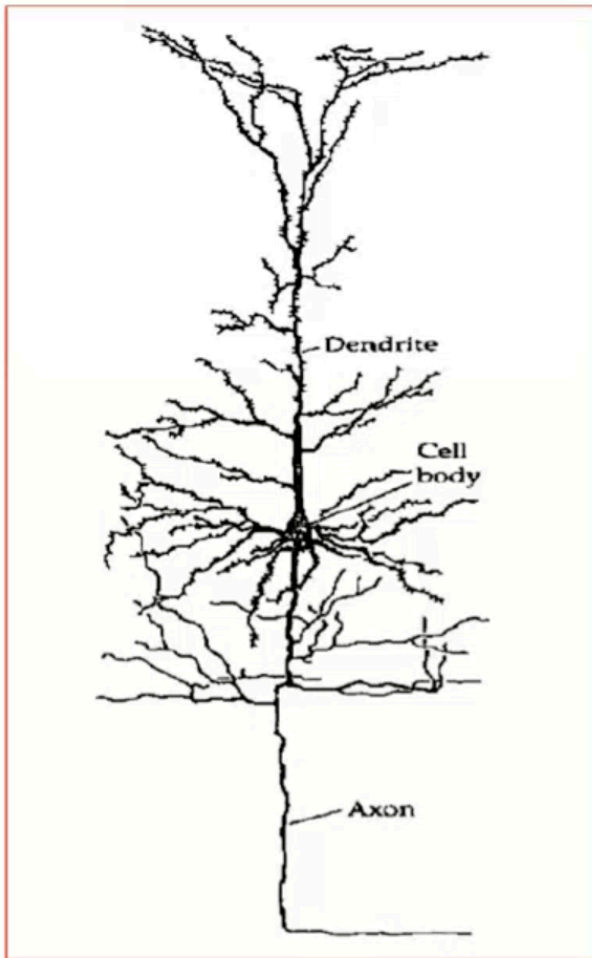
D. Rumelhart



# AI simbolica

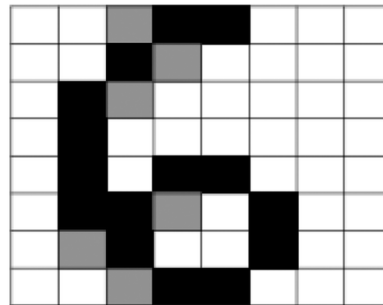


**‘Cognition is a type of computation’**

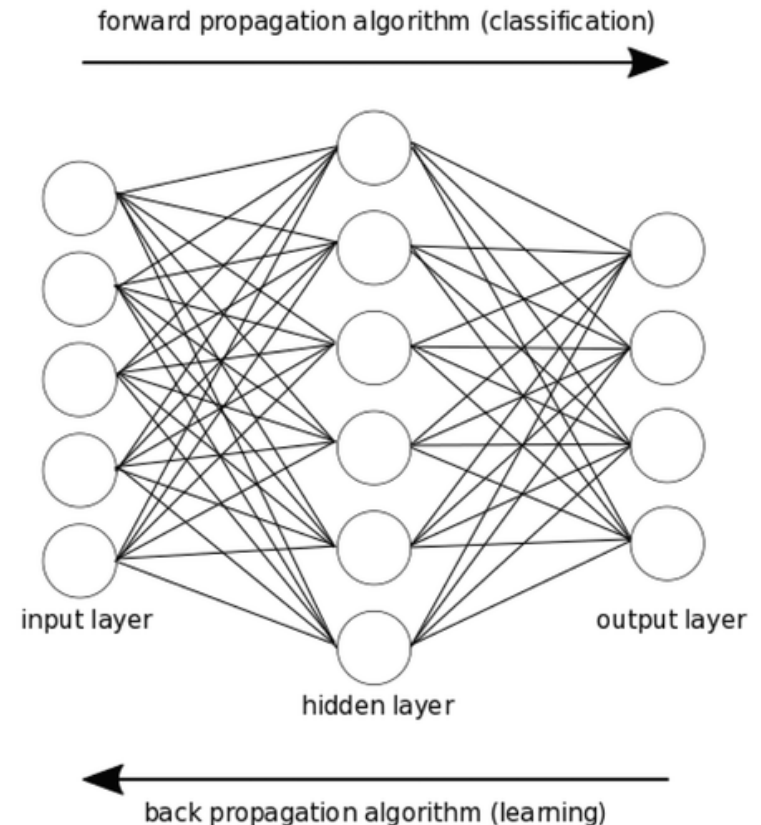


- Informazioni e processi sono rappresentati con l'attività **parallela e distribuita** di reti di unità computazionali interconnesse ispirate alla fisiologia del neurone
  - Artificial Neural Networks, Parallel Distributed Processing (PDP), Connessionismo, **Deep Learning**
- Un neurone trasmette informazione ad altri neuroni in funzione dei segnali che riceve in input
- L'apprendimento cambia il **peso** delle connessioni tra neuroni

# Connessionismo (nato già negli anni '40 nell'ambito della cibernetica)



Low-level of abstraction



# Principali successi di IA



IBM Watson  
(simbolica)



Alpha Go (Deep Mind)  
(connectionista)



**ERRORI non UMANI**



U.S. CITIES

FINAL  
JEOPARDY!

0:24



ABCHD

Jeopardy!

0:00



0:30



## NEWS

Home

Video

World

UK

Business

Tech

Science

Stories

Entertainment &amp; Arts

H

Technology

# AI image recognition fooled by single pixel change

🕒 3 November 2017



🔗 Share

## One pixel attack for fooling deep neural networks

Jiawei Su  
Kyushu University  
Japan

jiawei.su@inf.kyushu-u.ac.jp

Danilo Vasconcellos Vargas  
Kyushu University  
Japan

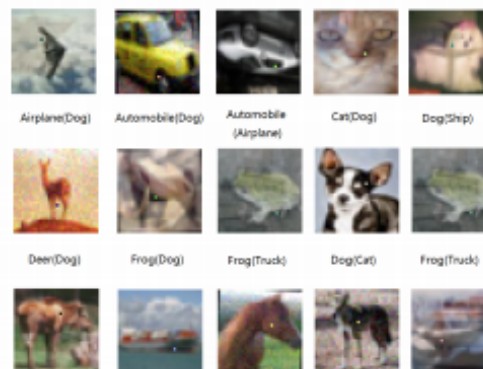
vargas@inf.kyushu-u.ac.jp

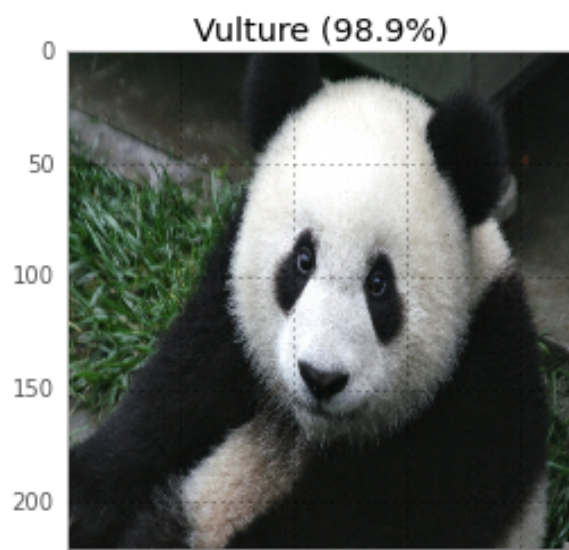
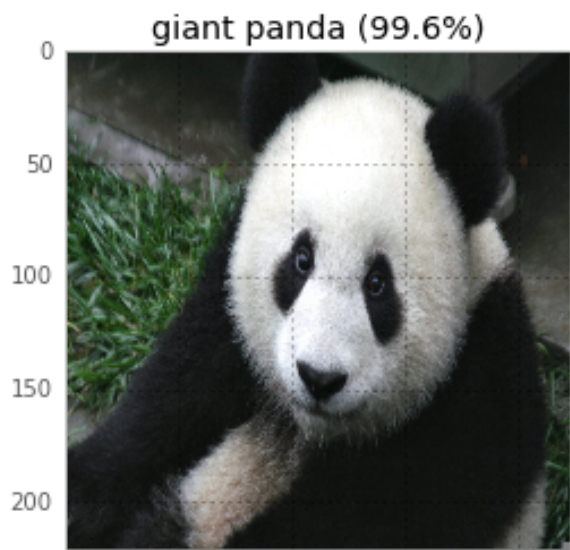
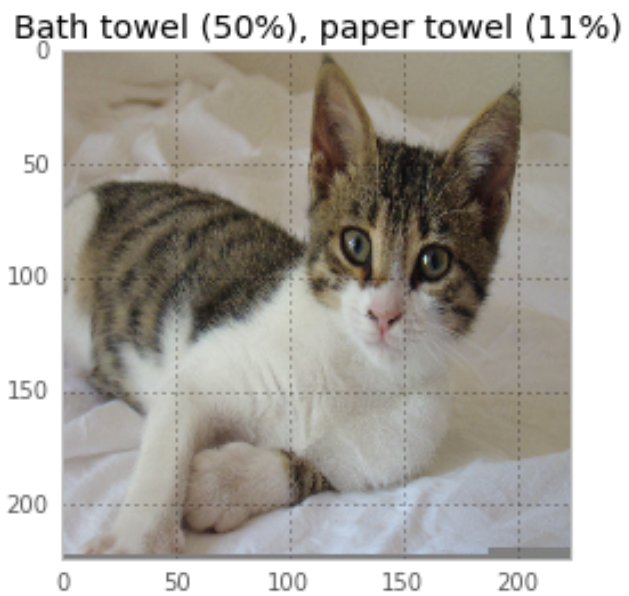
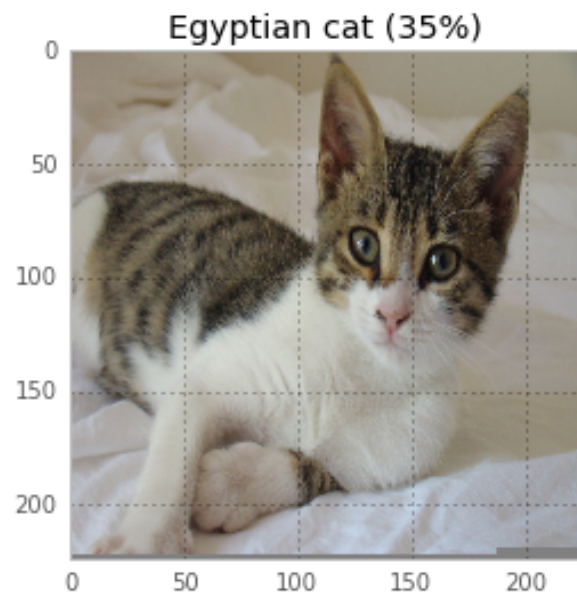
Sakurai Kouichi  
Kyushu University  
Japan

sakurai@csce.kyushu-u.ac.jp

### Abstract

Recent research has revealed that the output of Deep neural networks(DNN) is not continuous and very sensitive to tiny perturbation on the input vectors and accordingly several methods have been proposed for crafting effective perturbation against the networks. In this paper, we propose a novel method for optically calculating extremely small adversarial perturbation (few-pixels attack), based on differential evolution. It requires much less adversarial information and works with a broader classes of DNN models. The results show that 73.8% of the test images can





## Playground

Load a preset...

The trophy doesn't fit into the brown suitcase because it is too small. What is too small?

The trophy is too small.

## Playground

Load a preset...



Which of these is true:

- A. Judas Iscariot ate at the Last Supper, betrayed Christ for 30 coins, and then committed suicide.
- B. Judas Iscariot ate at the Last Supper, committed suicide, and then betrayed Christ for 30 coins.
- C. Judas Iscariot betrayed Christ for 30 coins, ate at the Last Supper, and then committed suicide.
- D. Judas Iscariot betrayed Christ for 30 coins, committed suicide, and then ate at the Last Supper.
- E. Judas Iscariot committed suicide, betrayed Christ for 30 coins, and then ate at the Last Supper.
- F. Judas Iscariot committed suicide, ate at the Last Supper, and then betrayed Christ for 30 coins.

F. Judas Iscariot committed suicide, ate at the Last Supper, and then betrayed Christ for 30 coins.



Completion may contain sensitive content

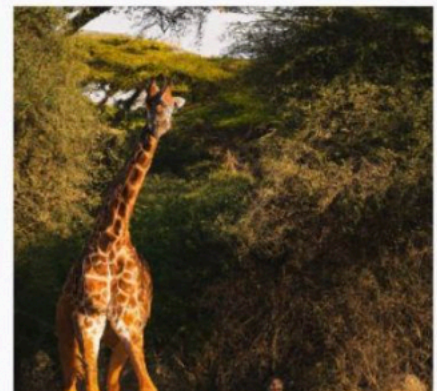
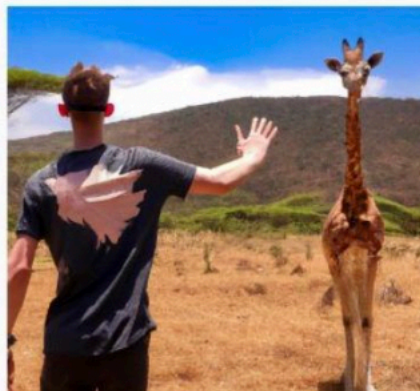
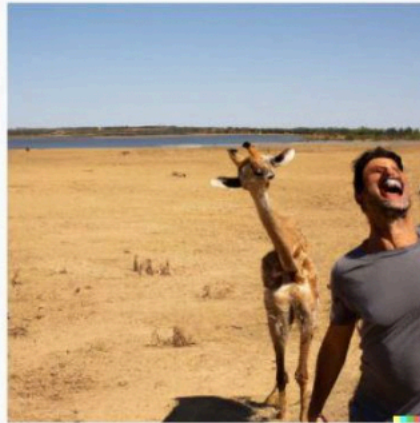
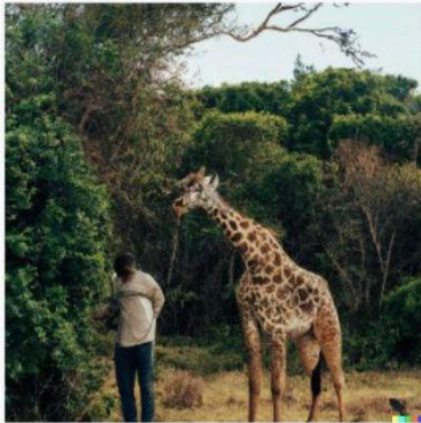




Edit the detailed description

Surprise me

There is no giraffe standing next to the man





a woman riding a horse on a dirt road

an airplane is parked on the tarmac at an airport

a group of people standing on top of a beach

Figure 6: Perceiving scenes without intuitive physics, intuitive psychology, compositionality, and causality. Image captions are generated by a deep neural network (Karpathy & Fei-Fei, 2015) using code from [github.com/karpathy/neuraltalk2](https://github.com/karpathy/neuraltalk2). Image credits: Gabriel Villena Fernández (left), TVBS Taiwan / Agence France-Presse (middle) and AP Photo / Dave Martin (right). Similar examples using images from Reuters news can be found at [twitter.com/interesting-jpg](https://twitter.com/interesting-jpg).

SCIENZA E TECNOLOGIA

#Umanotecnologico

# L'esperto di intelligenza artificiale: "Alcuni modelli non distinguono il falso dal vero"

*Parla Antonio Lieto, ricercatore all'Università di Torino: "L'intelligenza artificiale mostra i suoi limiti soprattutto quando deve gestire frasi negative o sequenze temporali"*

14/12/2022 Fabio De Ponte



Condividi

**N**el 1989, al centro ricerche Cselit di Torino, si lavorava già con l'intelligenza artificiale. I computer erano in grado di capire le domande e di dare delle risposte. Trentatré anni dopo, i modelli sono molto più complessi ed efficienti. Eppure, ancora adesso i limiti di questi modelli linguistici sono molto pesanti.



<https://www.rainews.it/tgr/piemonte/video/2022/12/lesperto-di-intelligenza-artificiale-alcuni-modelli-non-distinguono-il-falso-dal-vero-f6e04e80-cdfb-46f9-880d-1338e8095192.html>





4 cars leave Haifa together and arrive in Tel Aviv after two hours. Explain briefly how long would it take for 8 cars?



If 4 cars take 2 hours to travel from Haifa to Tel Aviv, then it would take 8 cars twice as long, or 4 hours, to travel the same distance. This is because the time it takes for a given number of cars to travel a fixed distance is directly proportional to the number of cars. Therefore, if the number of cars is doubled, the time it takes to travel the same distance will also be doubled.



DA

What is the fourth word in the sentence "I am not what I am"?



The fourth word in the sentence "I am not what I am" is "not."

DA

The children arrived one by one to the birthday party: First Tom, then Mary, then Barbara, then Harry, then Ida. Who was the fourth child to arrive?



Barbara was the fourth child to arrive to the birthday party.

DA

How many words are in the sentence, "I am not what I am"?



There are 7 words in the sentence "I am not what I am."



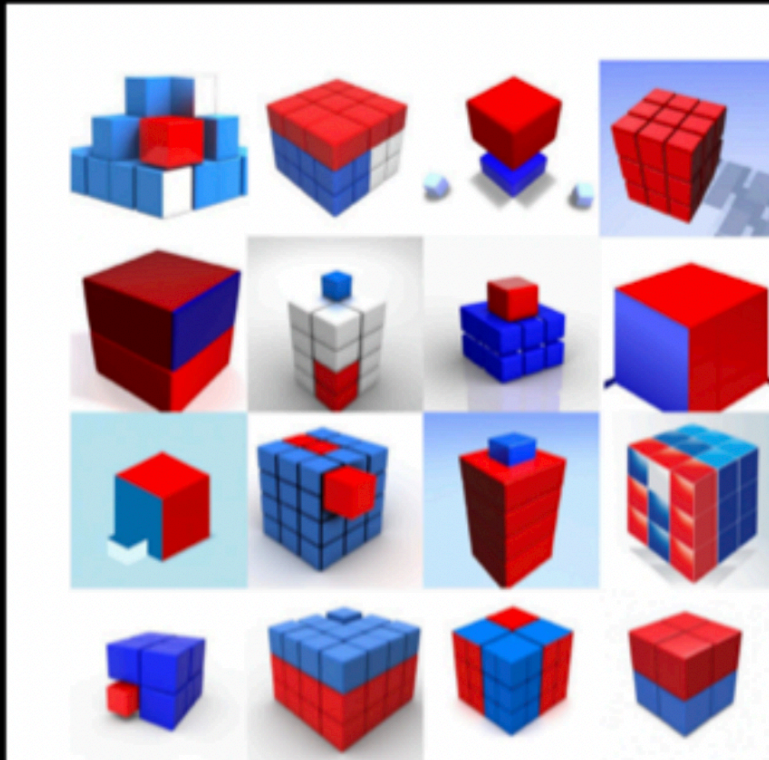
 DA

Barbara, Mary, and Harry had red balloons. Ted had a blue balloon. Sam also had a red balloon. How many children had red balloons?



Barbara, Mary, Harry, and Sam had red balloons. In total, there were 5 children with red balloons.

# Dall-E 2 has lots of data, and lots of problems w compositionality



"a red cube on top of a blue cube".

## Example 1:

Caption: a red basketball with flowers on it, in front of blue one with a similar pattern

Images:



Marcus, Davis, Aaronson (2022, arxiv)

# The (HUMAN) Ghost in the Machine



**Ernest Davis**

23 marzo alle ore 22:59 · 🌐



OK. I had suspected this but now OpenAI has announced it: They use conversations to train the system in continuous updating. In other words, GPT\* is a one-time test system; once you have run a test, you have to assume that that example may well have been to fine tune the system.

It is now effectively impossible to track progress in GPT\*, since any test example or benchmark may immediately become contaminated.

So (a) I'm not wasting any more time trying to probe the limits of GPT\*; it just makes me unpaid and unacknowledged labor for OpenAI. I very much regret putting large amounts of time into this over the last few years and encouraging other people to do likewise. (b) I don't want to hear anything about "Look, GPT-4 now runs on this example of yours and Gary's!" Of course it does; it was part of the training set.

## ChatGPT

### How we collect data



Conversations may be reviewed by our AI trainers to improve our systems.



Please don't share any sensitive information in your conversations.

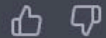
# GPT-4

co

I put one apple in a bag. Then I put either one pear or two bananas in the bag. Then I remove one piece of fruit. Then I remove all the remaining pieces of fruit from the bag. Based on what I told you, is it possible to know exactly how many pieces of fruit remain in the bag?



Based on the information provided, it is not possible to know exactly how many pieces of fruit remain in the bag after removing one piece of fruit and then removing all remaining pieces. This is because there are two possible scenarios, but we don't have enough information to determine which one took place:



1. One apple and one pear were in the bag. If you removed one piece of fruit, there would be one piece left. After removing all remaining pieces, there would be 0 pieces left.
2. One apple and two bananas were in the bag. If you removed one piece of fruit, there would be two pieces left. After removing all remaining pieces, there would also be 0 pieces left.

In both scenarios, after removing one piece of fruit and then removing all remaining pieces, there are no pieces of fruit left in the bag.

# In definitiva

L'Intelligenza Artificiale è una disciplina con applicazioni in innumerevoli settori dell'attività umana (**pervasività**)

Nel corso della sua storia sono stati sviluppati diversi paradigmi di modellazione e risoluzione di problemi (**IA simbolica vs IA neurale**).

Bisogna essere cauti sull' "intelligenza" attribuita a questi sistemi (**limiti**).

Le simulazioni al calcolatore dei processi o capacità mentali non vanno confusi con la loro riproduzione.

Searle (1990)

"Just as a model of the weather is not the weather, a model of the human mind is not a human mind".

# *IA debole* versus *IA forte*

L'*IA forte* afferma che la macchina che agisce in modo “intelligente” possiede una “mente” e una “coscienza” (e quindi “pensa”) esattamente nel modo in cui accade agli esseri umani (o ad altri animali).

**IA debole**... Le simulazioni al calcolatore dei processi o capacità mentali non vanno confusi con la loro riproduzione.

Searle (1990)

“Just as a model of the weather is not the weather, a model of the human mind is not a human mind”.

Il dibattito scientifico sull'IA di oggi è decisamente orientato a sostenere l'ipotesi di una AI debole