Conceptualizing Online Brain Manipulation

Claude Castelluccia

Claude.castelluccia@inria.fr

2023

From Dataveillance to Online Datapulation

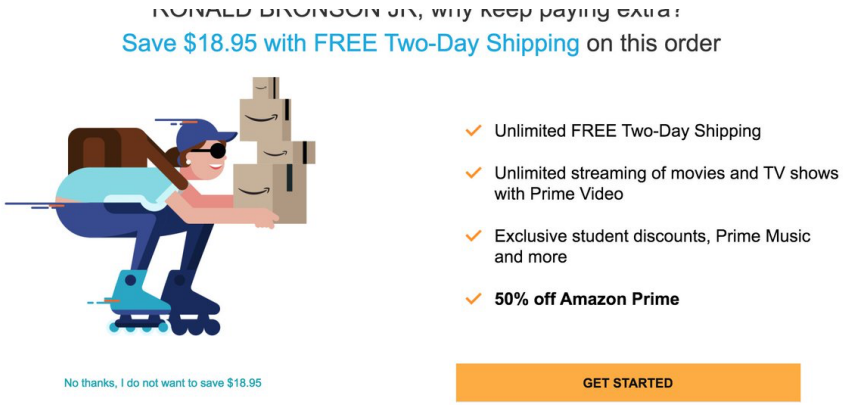
- Today the smallest details of our daily lives are tracked and traced more closely than ever before (**liquid** surveillance or **Dataveillance**)!

- However, data is not only used for surveillance anymore… **but also to manipulate people's opinions and decisions!**
  - Fake News
  - Filtering, censorship,
  - Psychological profiling: Stories around Trump election, Brexit,...

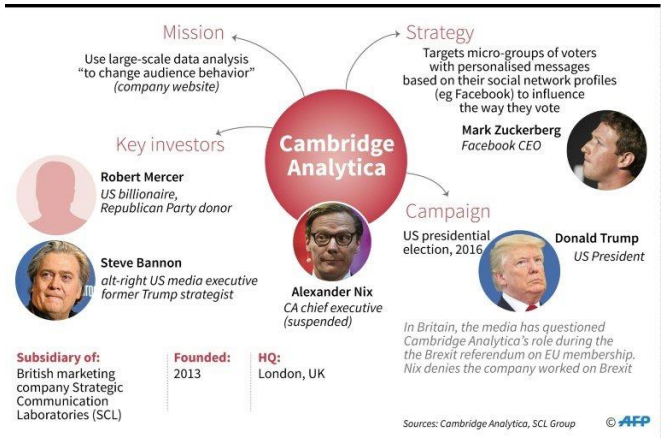
- Datapulation: Data + Manipulation

Different types of Online Manipulation

Fake News



Dark Patterns



Personalized Manipulation (Cambridge Analytica)

The need for Conceptualization

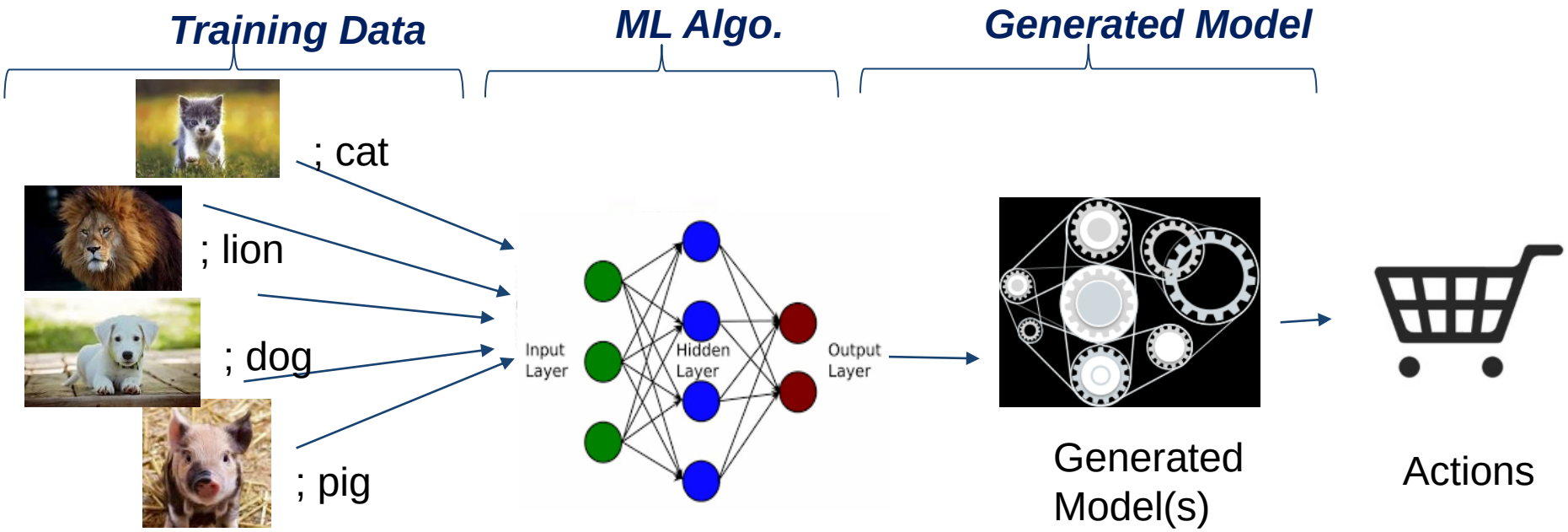
- Data manipulation (and brain insecurity) is growing
- We need to define technical solutions and regulate t!
- But policy makers have struggled to respond since the issue is under-conceptualized!
- Our approach:
  - we get inspired by CS security to conceptualize "security" properties
  - We get inspired by AI security to conceptualize attack vectors

Brain Online Manipulation
Security Properties

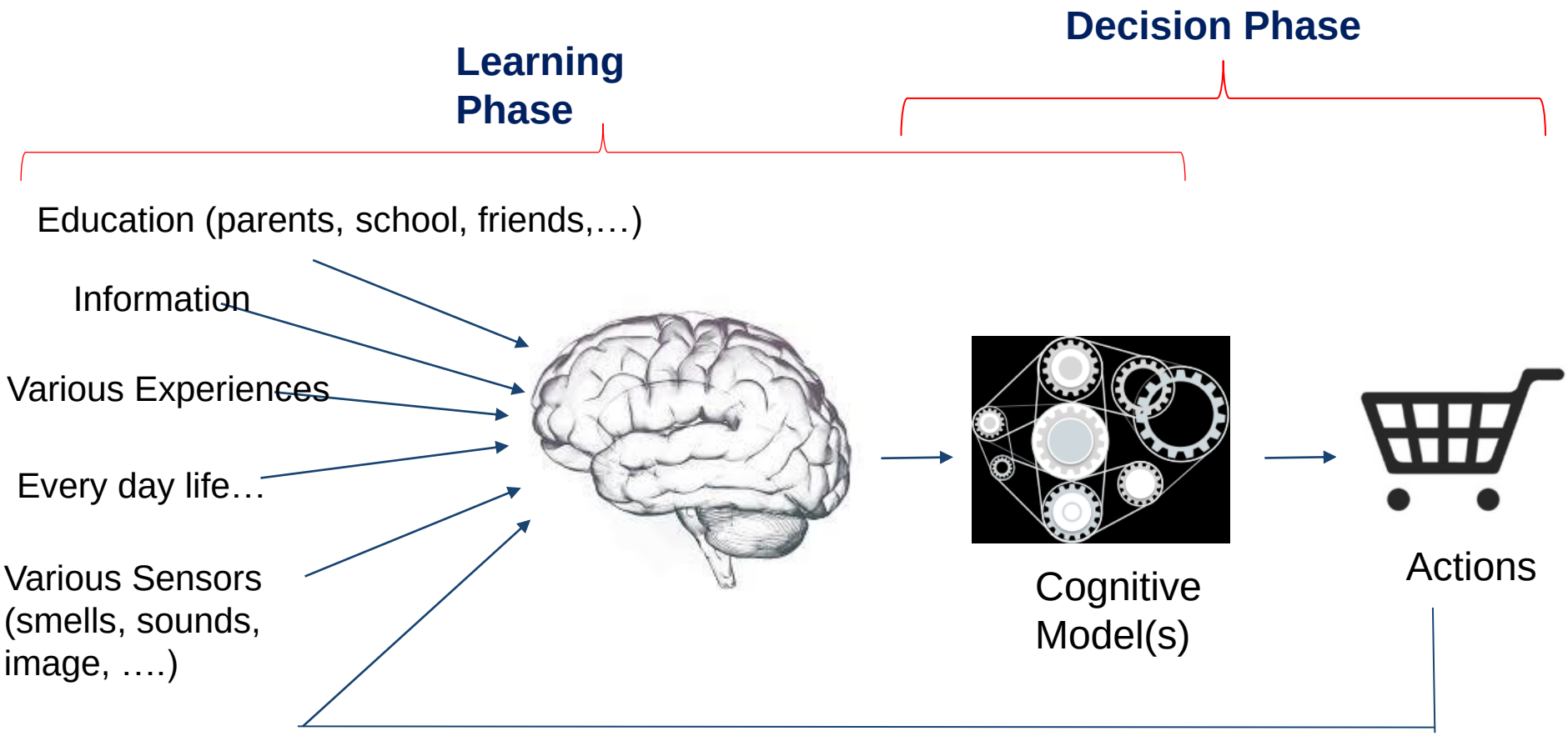
- The brain can be considered as a Black box
- What do it mean to secure a black box (CIA)?
  - **C**onfidentiality
  - **I**ntegrity
  - **A**vailability

  } *Data Manipulation*
- Confidentially (Mental liberty) is also very concerning
  - See recent dev. combining AI + EEG to "read" the mind

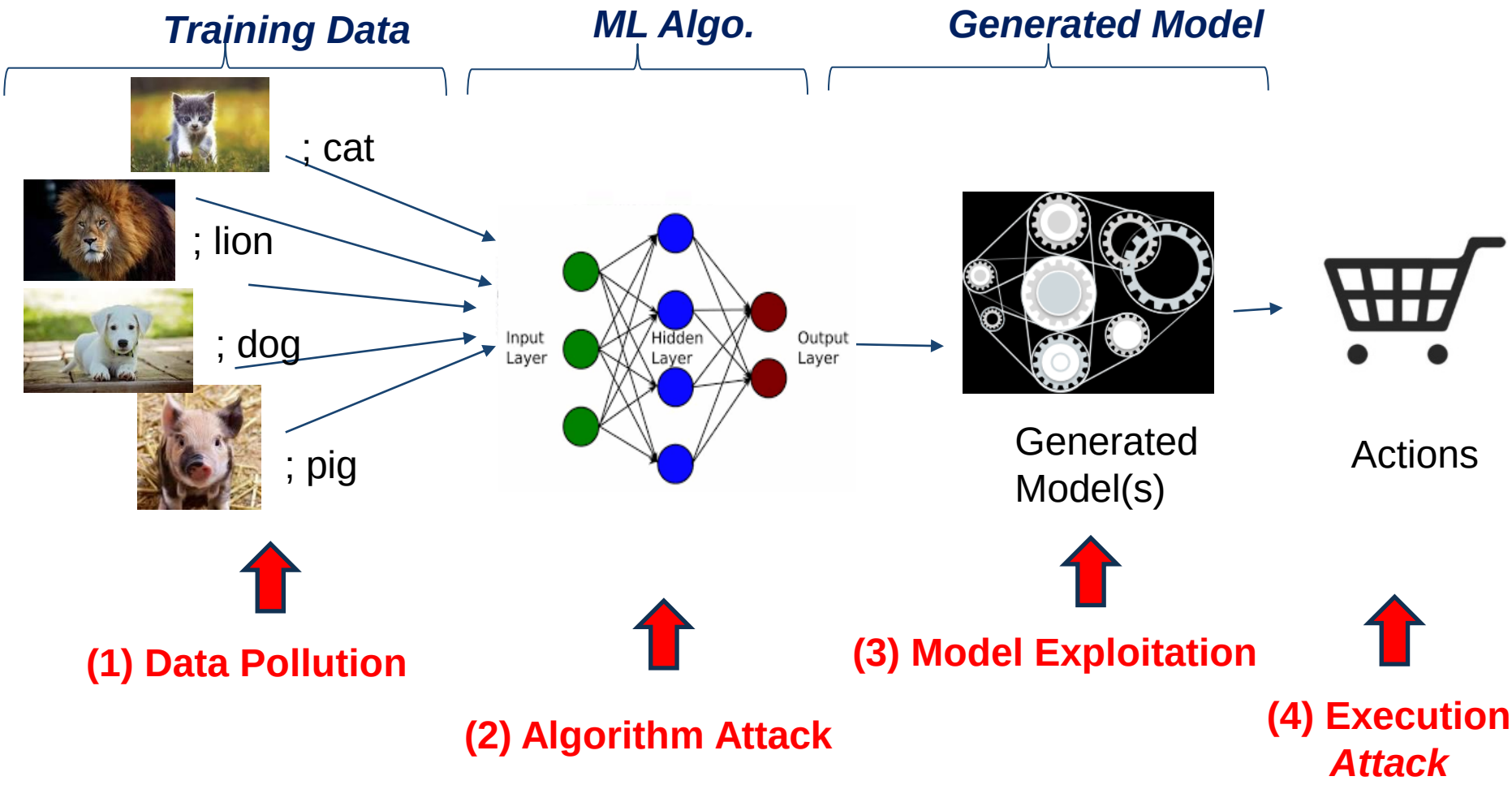
Brain Manipulation Attack Vectors
(getting inspired from Machine Learning)



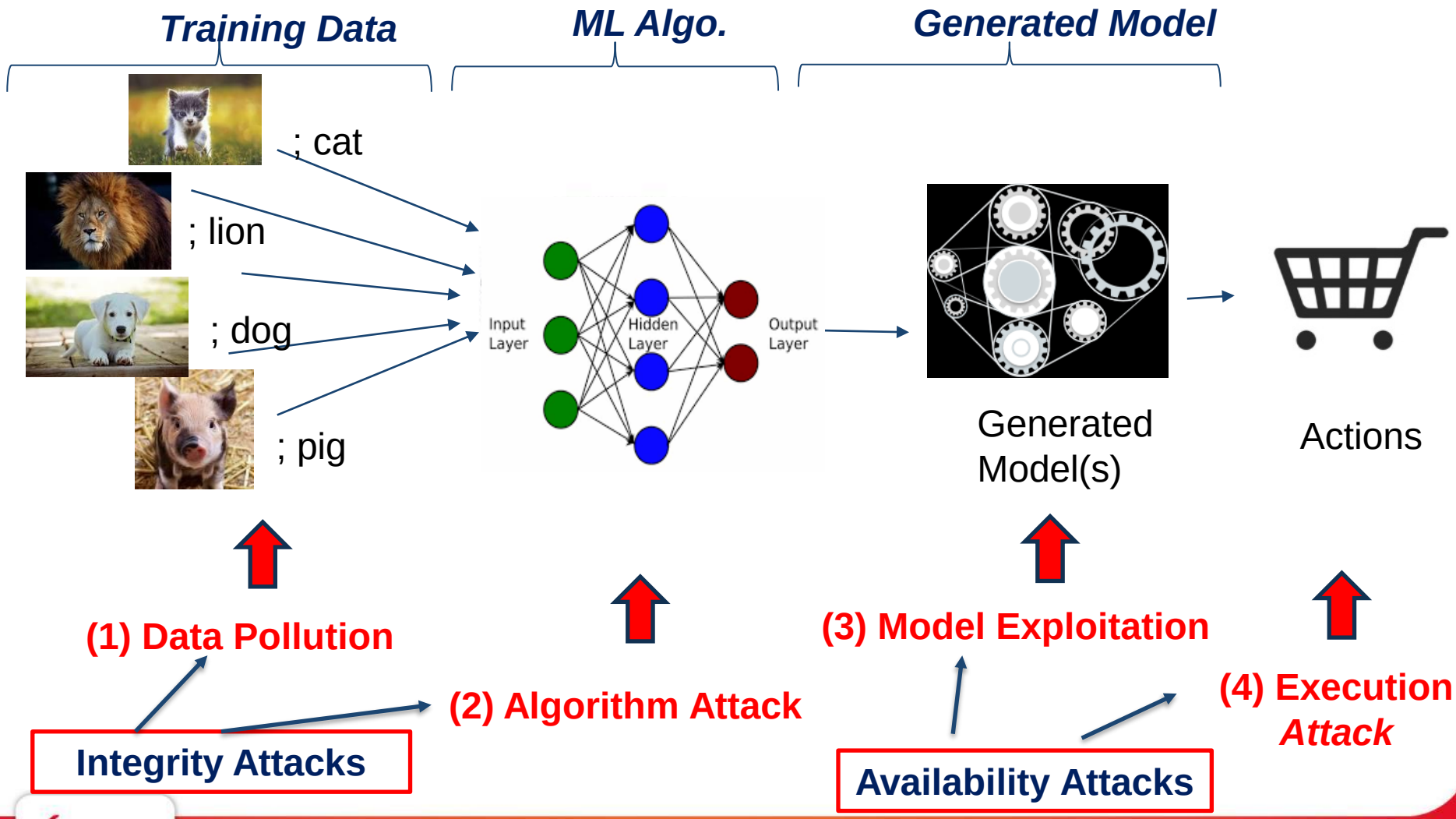
How do People Learn and make Decisions?
A (very) Simple Meta-Model



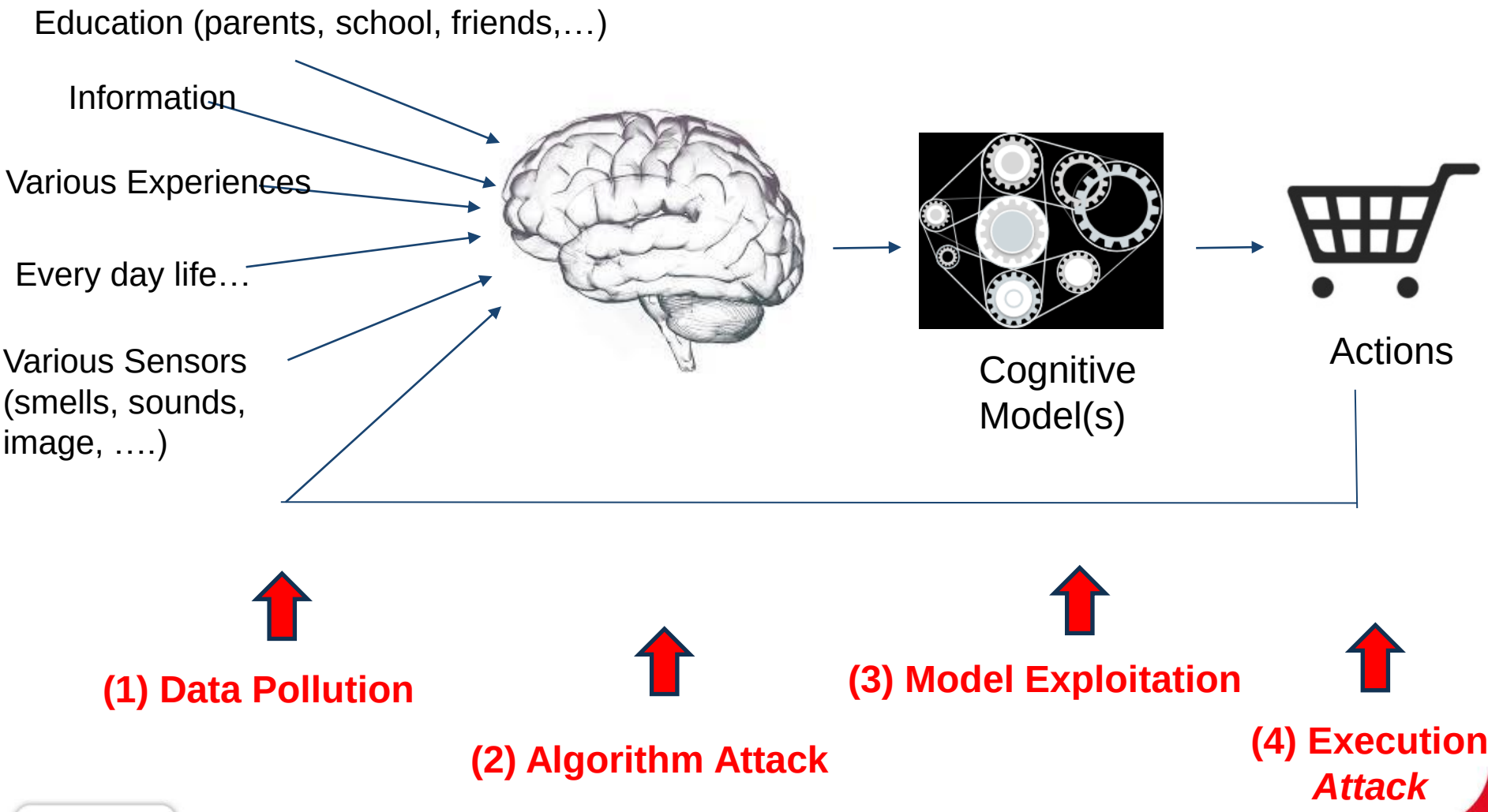
Machine Learning (In) Security
Attack vectors



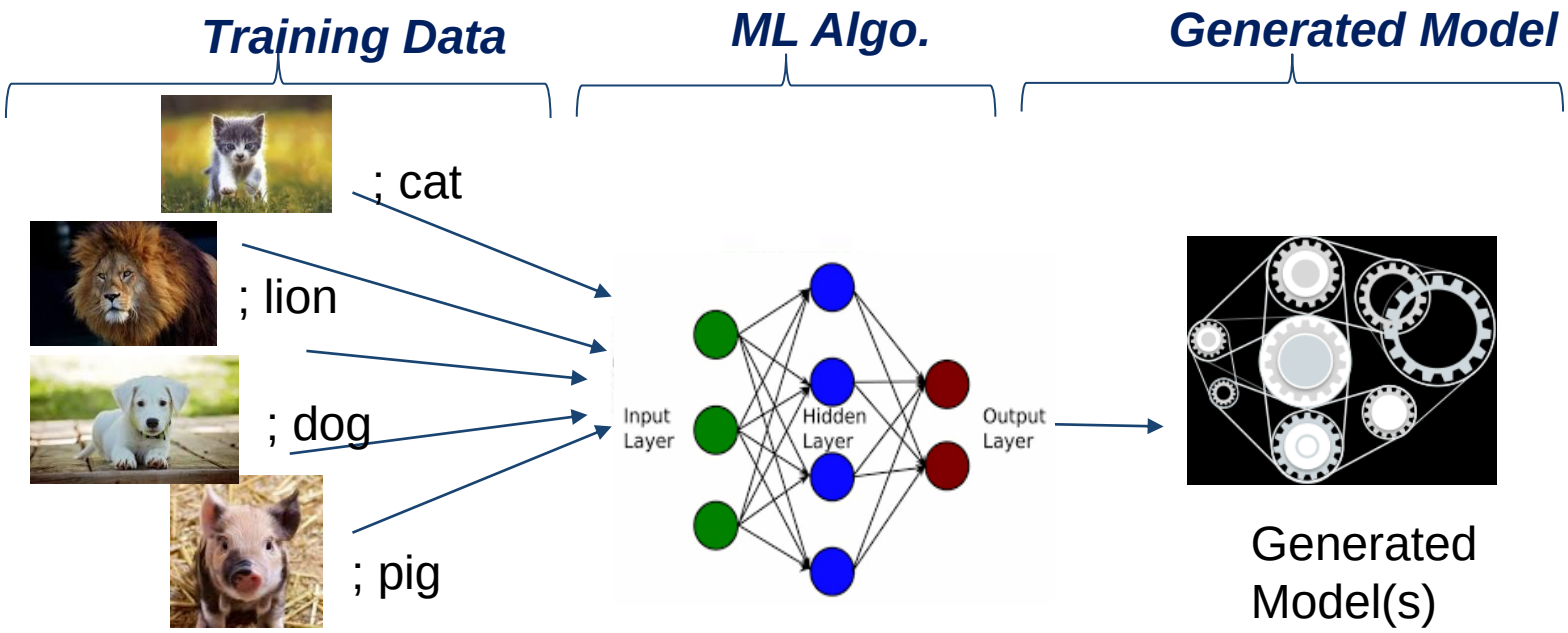
Machine Learning (In) Security
Attack vectors



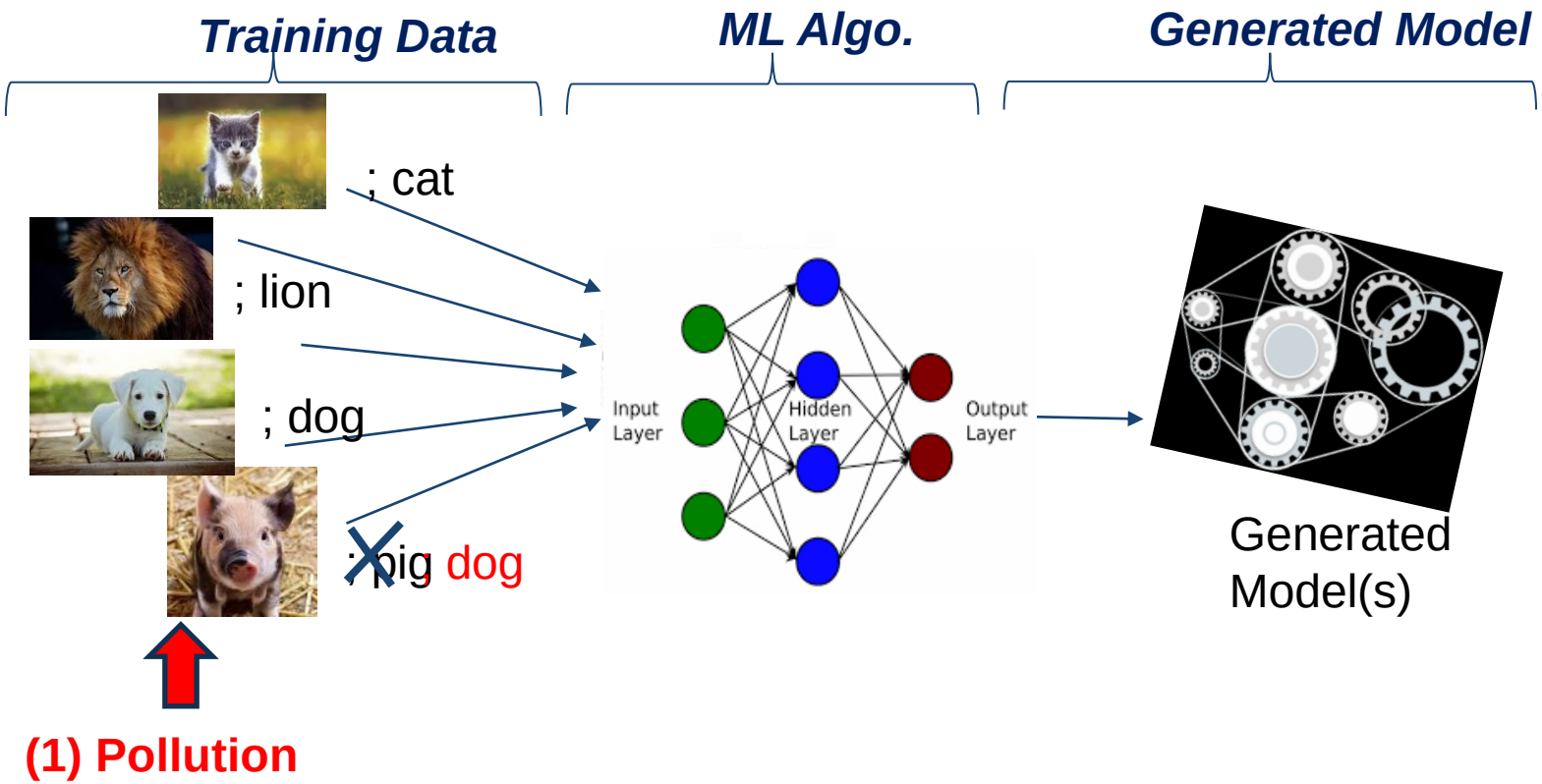
How to Manipulate Human?



Similar to Pollution Attacks in Machine Learning



Machine Learning & Pollution Attacks



Brain Pollution Attack (Integrity Attack)

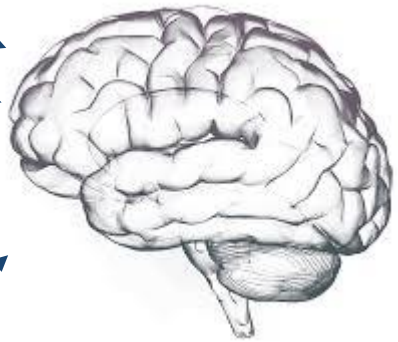
Education (parents, school, friends,…)

Information

Various Experiences

Every day life…

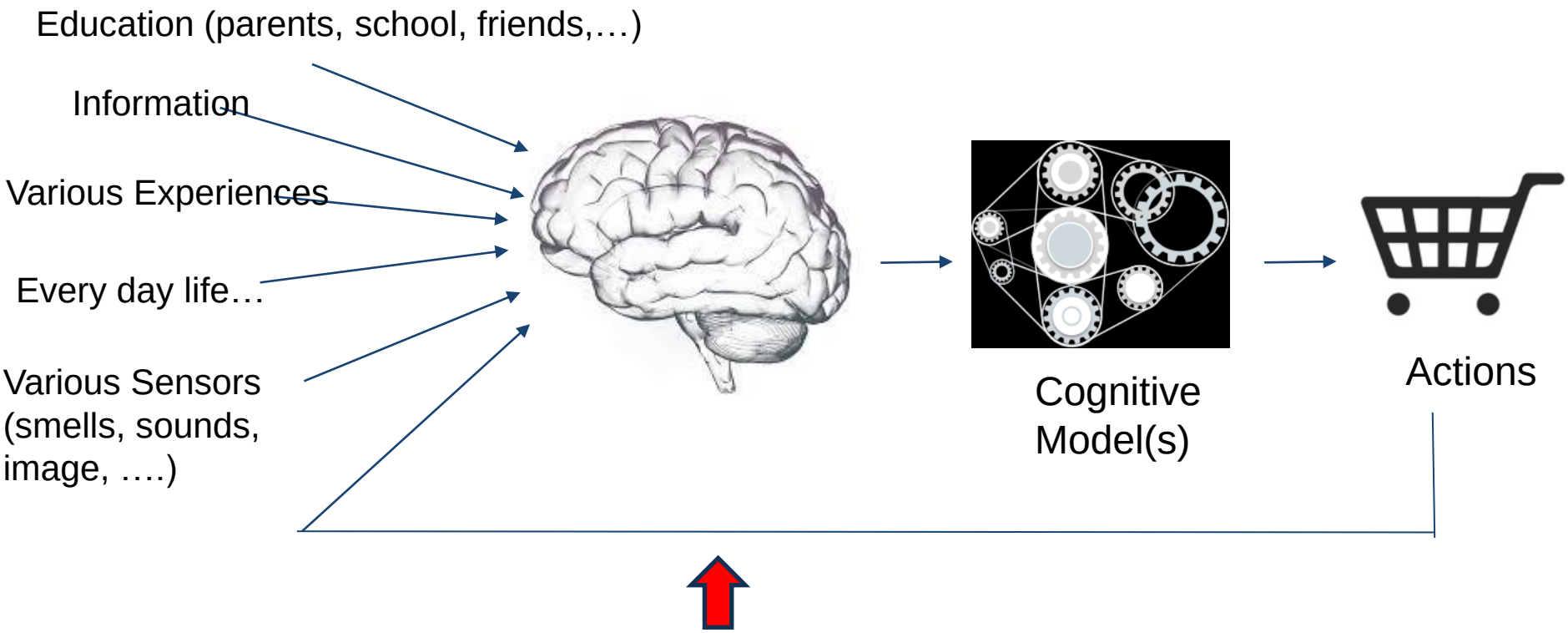
Various Sensors (smells, sounds, image, ….)



Cognitive Model(s)

Actions

How to Manipulate Human?
Physical Attacks

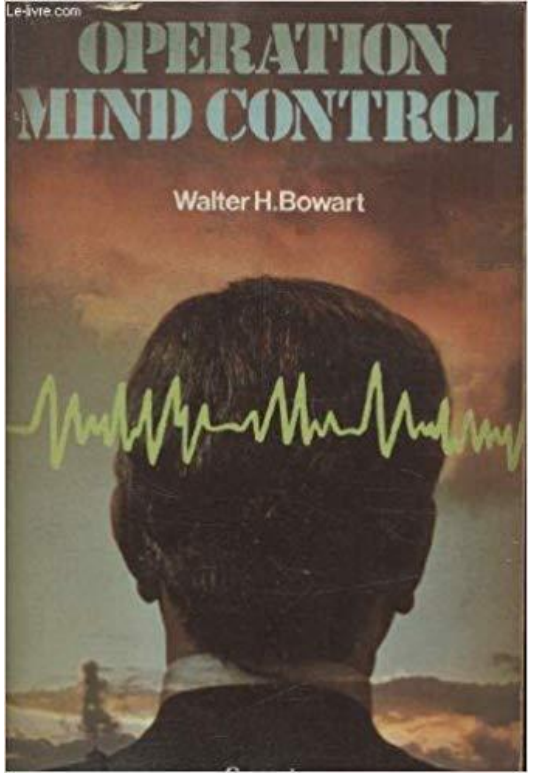


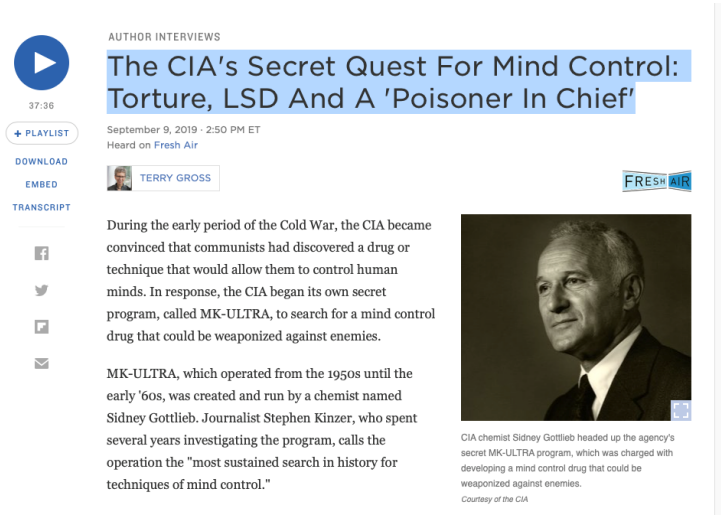
(2) Physical Attacks:
Attacks that modify its
Operation (physical attack, brain washing, violence,
Drug injection, physical threats!

In Search of Mind Control

Various « *brain-washing* » techniques, tested by CIA, based on drugs (LSD), hypnosis, electronic brain stimulation, low-frequency sounds …

Ex: CIA's MK-ULTRA program in the 50s that searches for a mind control drug that could be weaponized against enemies.





37:36

+ PLAYLIST

DOWNLOAD

EMBED

TRANSCRIPT

AUTHOR INTERVIEWS

The CIA's Secret Quest For Mind Control: Torture, LSD And A 'Poisoner In Chief'

September 9, 2019 · 2:50 PM ET

Heard on Fresh Air

TERRY GROSS

FRESH AIR

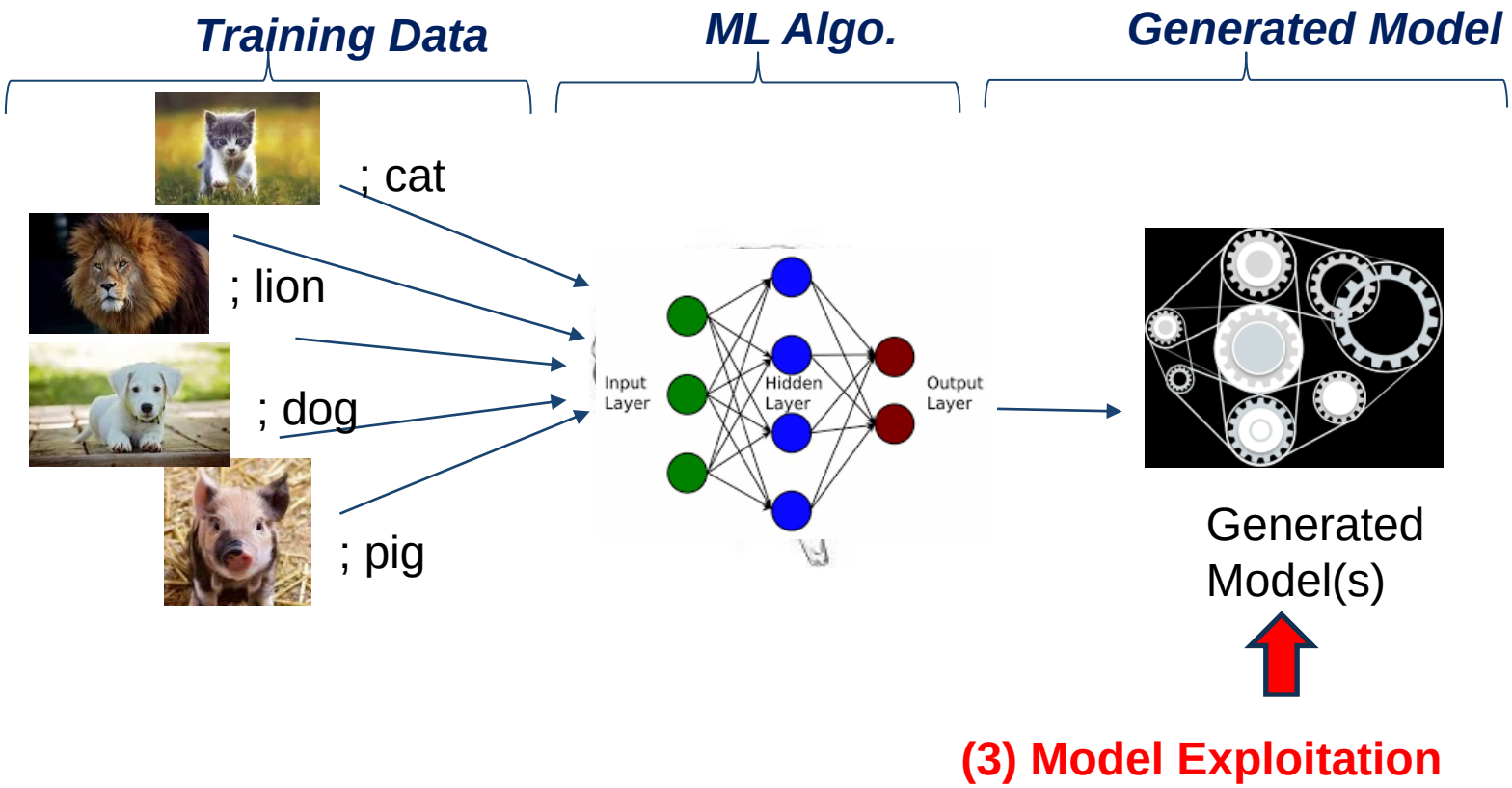
During the early period of the Cold War, the CIA became convinced that communists had discovered a drug or technique that would allow them to control human minds. In response, the CIA began its own secret program, called MK-ULTRA, to search for a mind control drug that could be weaponized against enemies.

MK-ULTRA, which operated from the 1950s until the early '60s, was created and run by a chemist named Sidney Gottlieb. Journalist Stephen Kinzer, who spent several years investigating the program, calls the operation the "most sustained search in history for techniques of mind control."

CIA chemist Sidney Gottlieb headed up the agency's secret MK-ULTRA program, which was charged with developing a mind control drug that could be weaponized against enemies.

Courtesy of the CIA

Machine Learning & Adversarial Examples (Availability Attack)



Machine Learning Adversarial Example

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

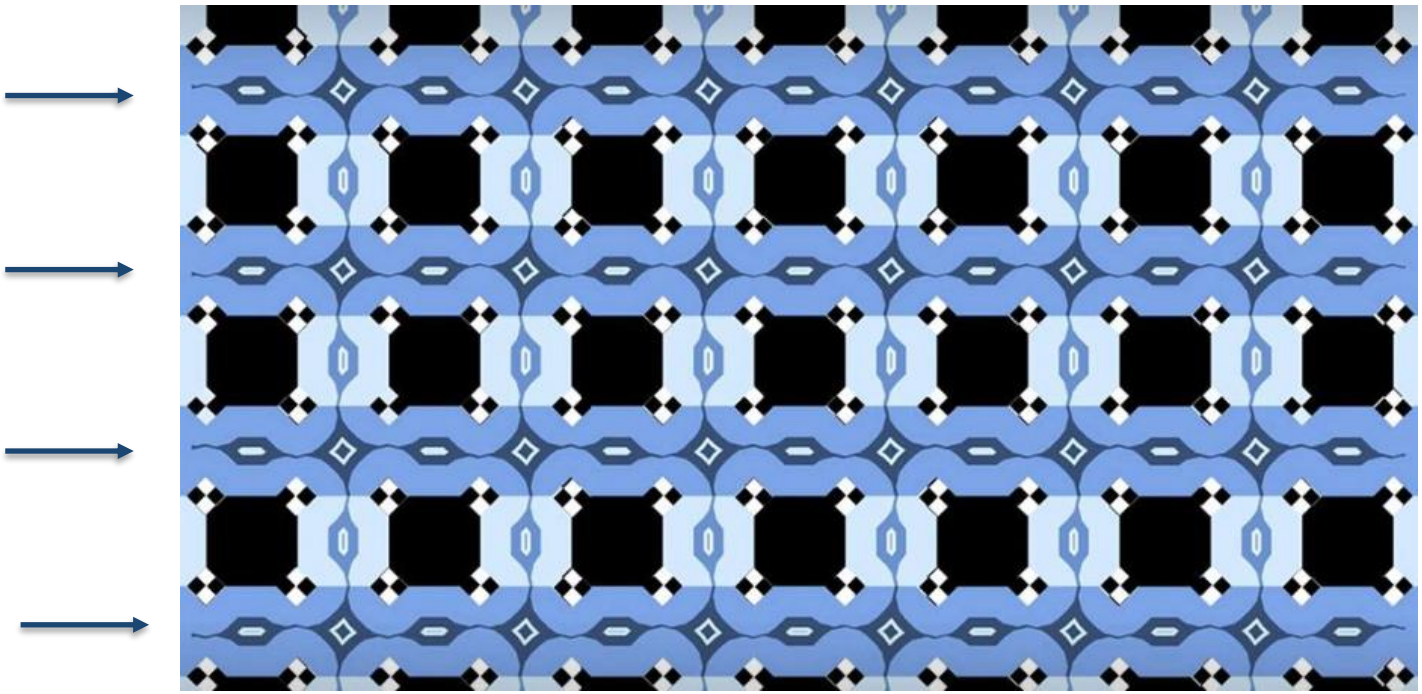
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

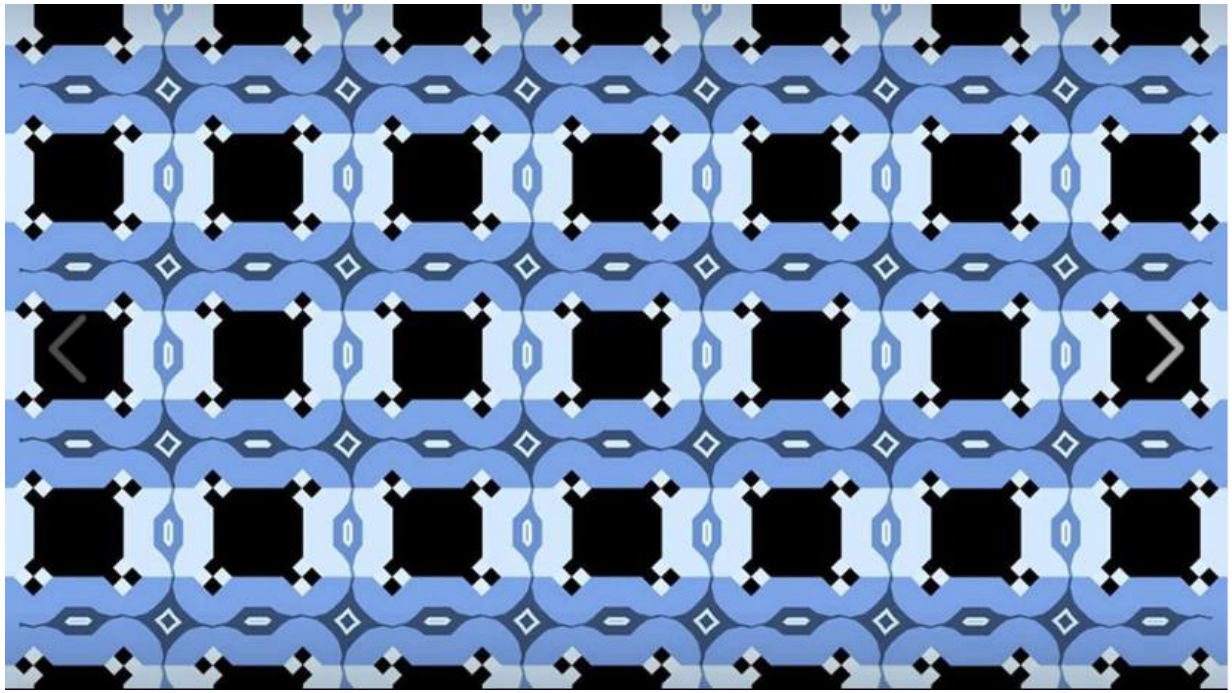
Cognitive Adversarial Example

- Those lines are parallel!



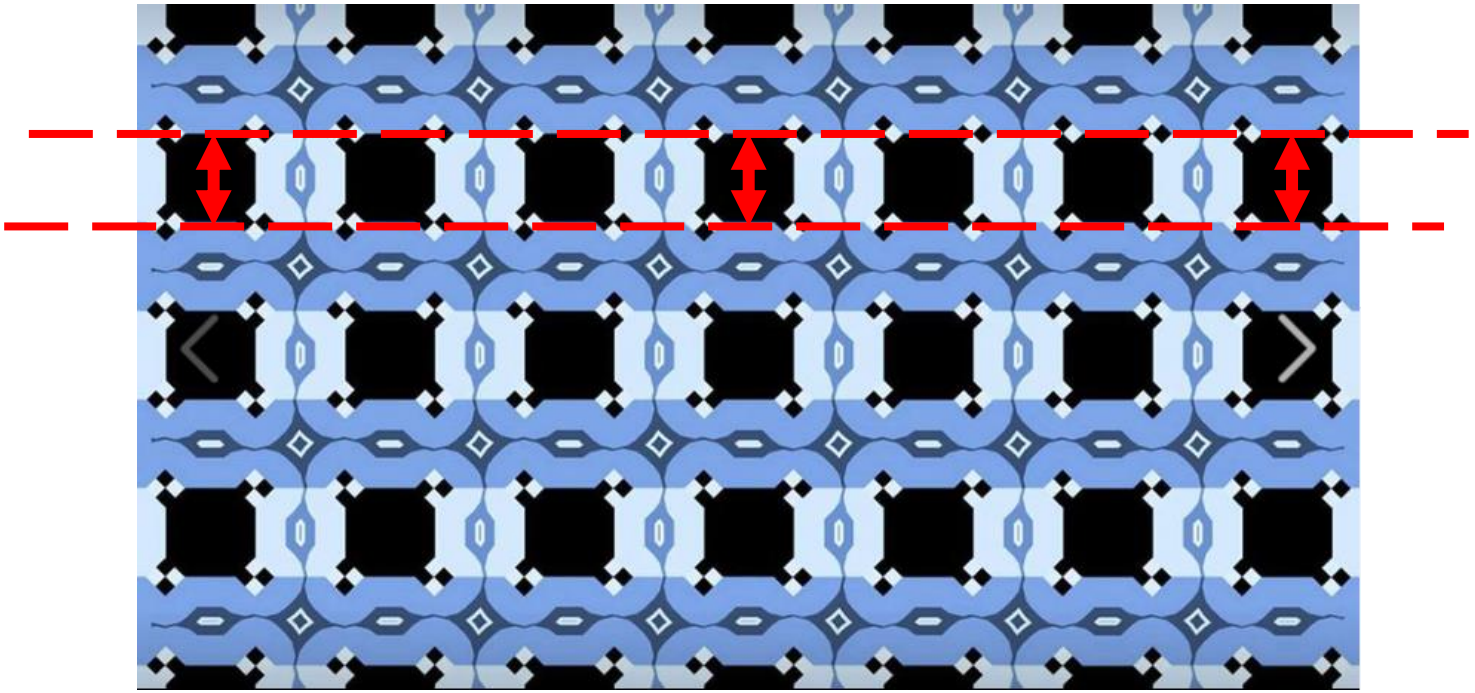
And now?

- Those lines are still parallel!



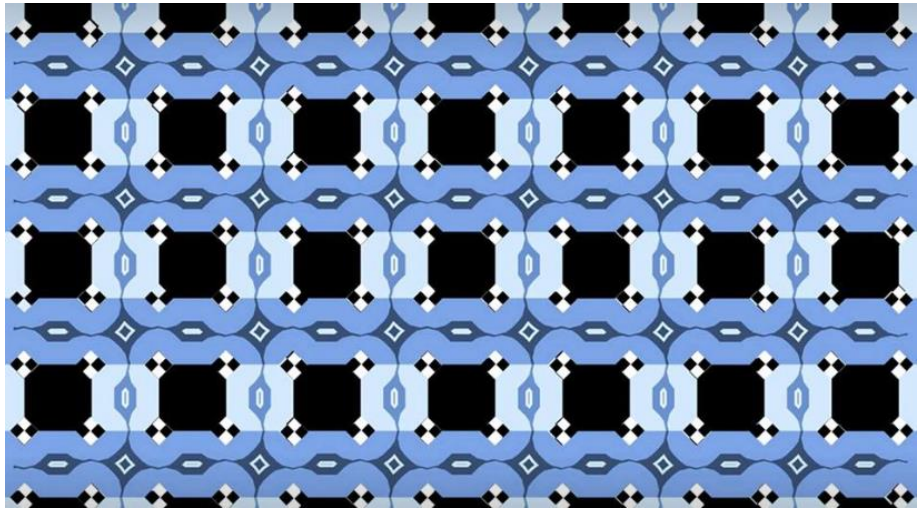
And now?

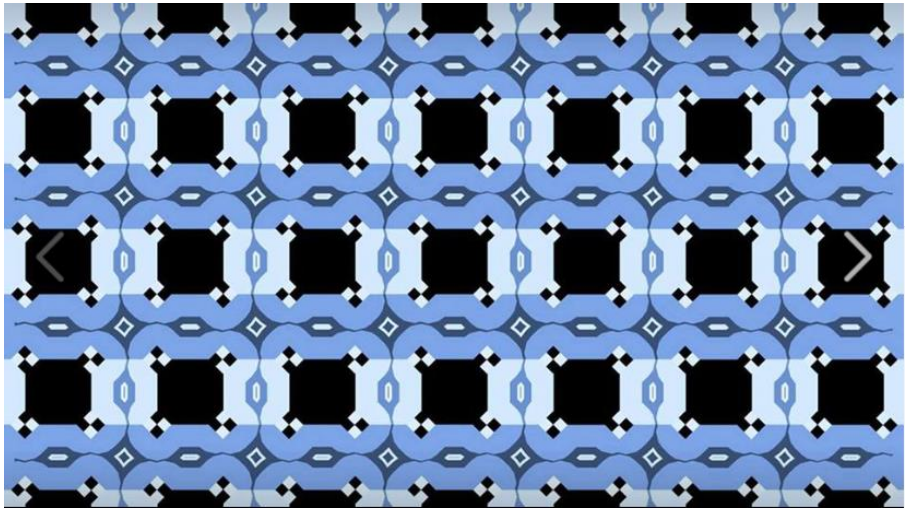
- Those lines are still parallel (cognitive dissonance)!



Cognitive Adversarial Example

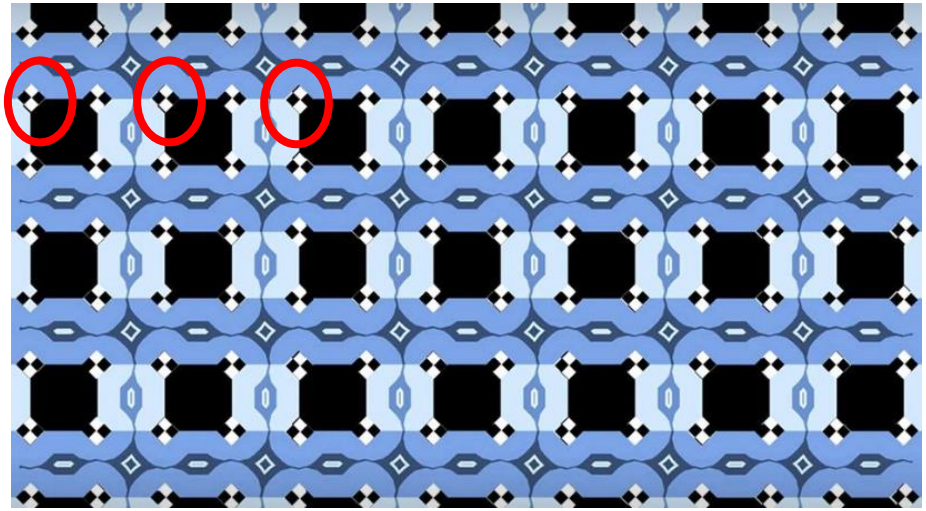
- What happened?

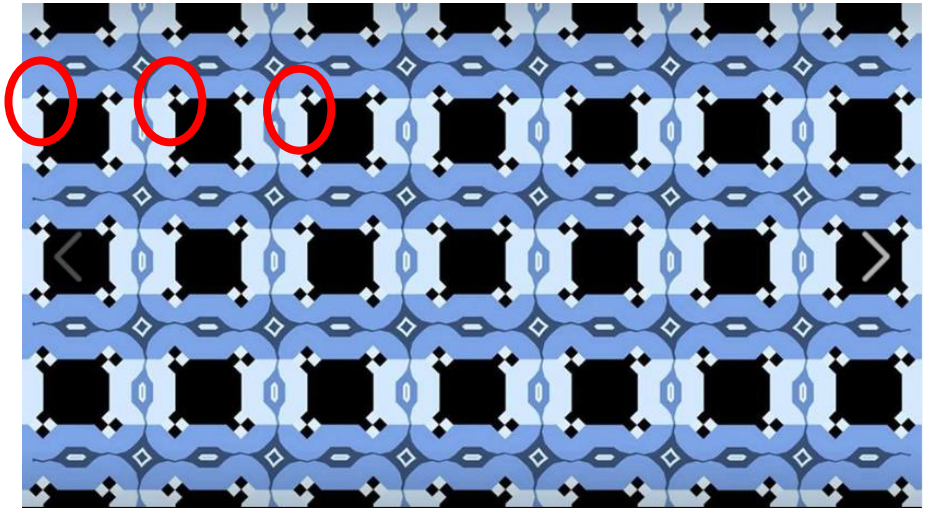




Cognitive Adversarial Example

- What happened?

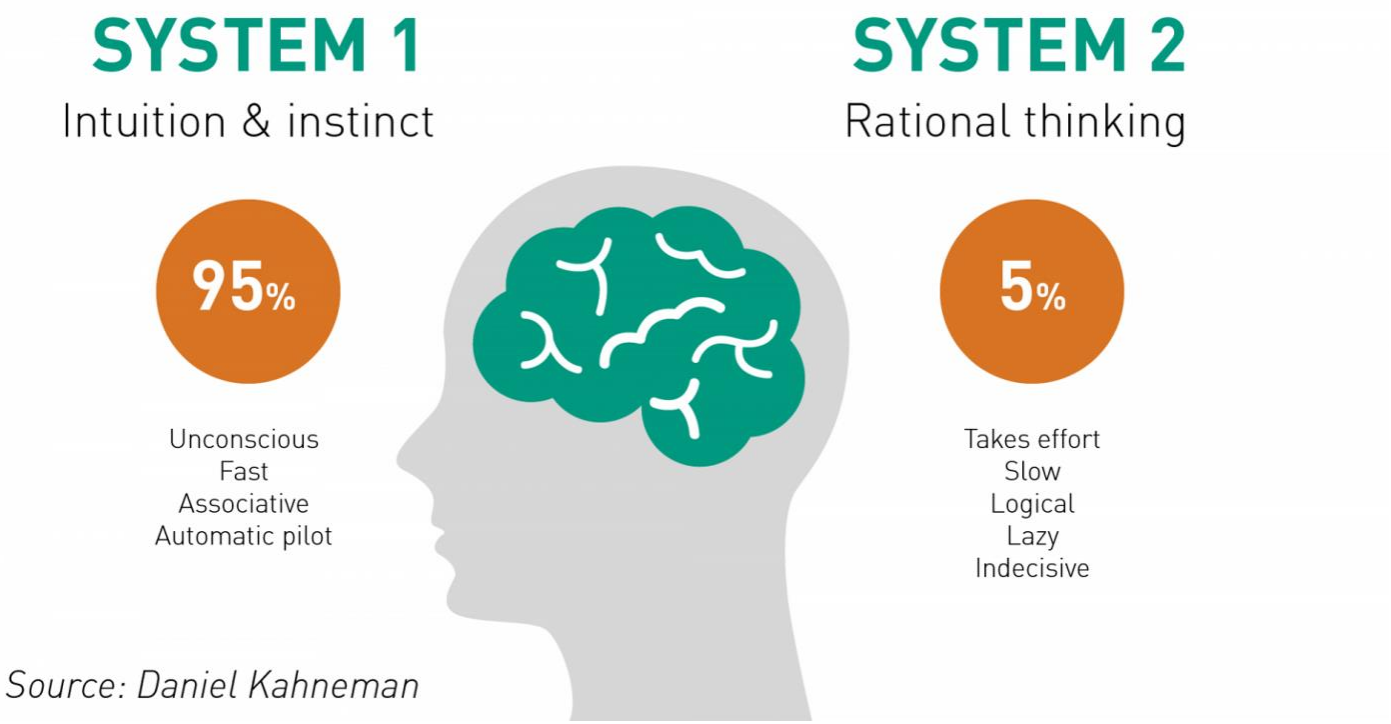




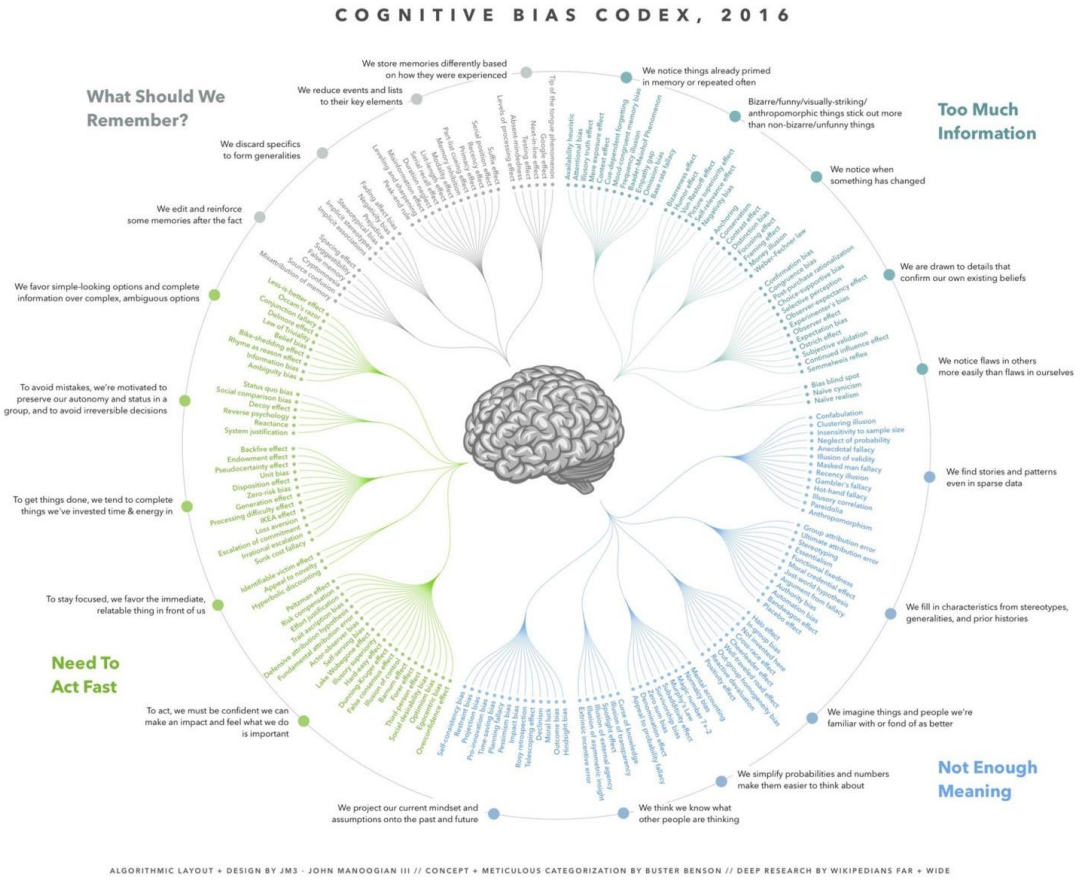
Other Manipulations based on Cognitive Biases

- **Emotion Manipulation**
  - Emotions, such as anger, fear, sadness, are contagious
  - A Facebook controversial study showed that emotion propagation can happen online, without direct interaction, and can be manipulated, for example, by propagating fake news or “likes”.
  - Concerning since emotions influence our decisions

Most decisions are based on Cognitive Biases

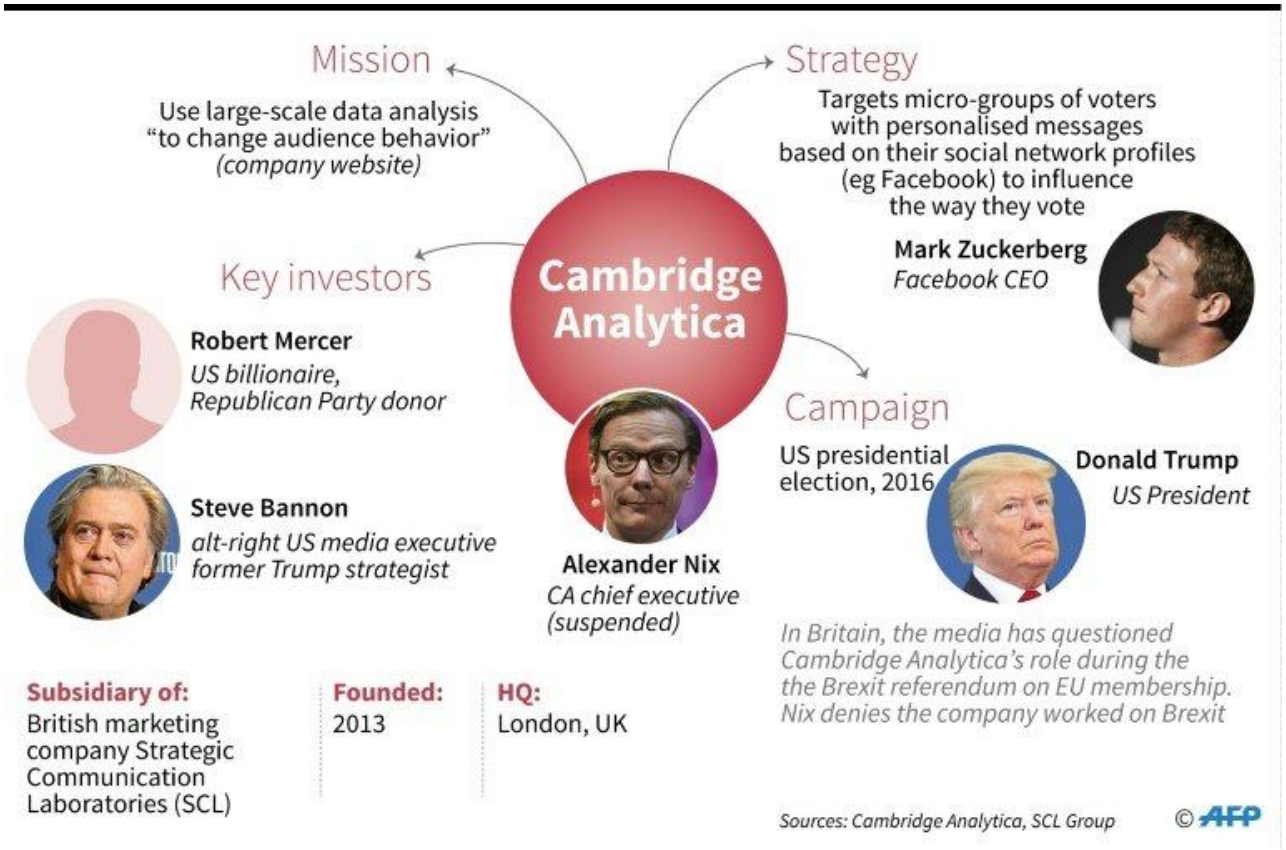


Bad News: There are many cognitive biaises!



Targeted Cognitive Bias Exploitation

- Targeted Manipulation is even more powerful!
- Cambridge Analytica has shown that psychological **online** profiling is easy!



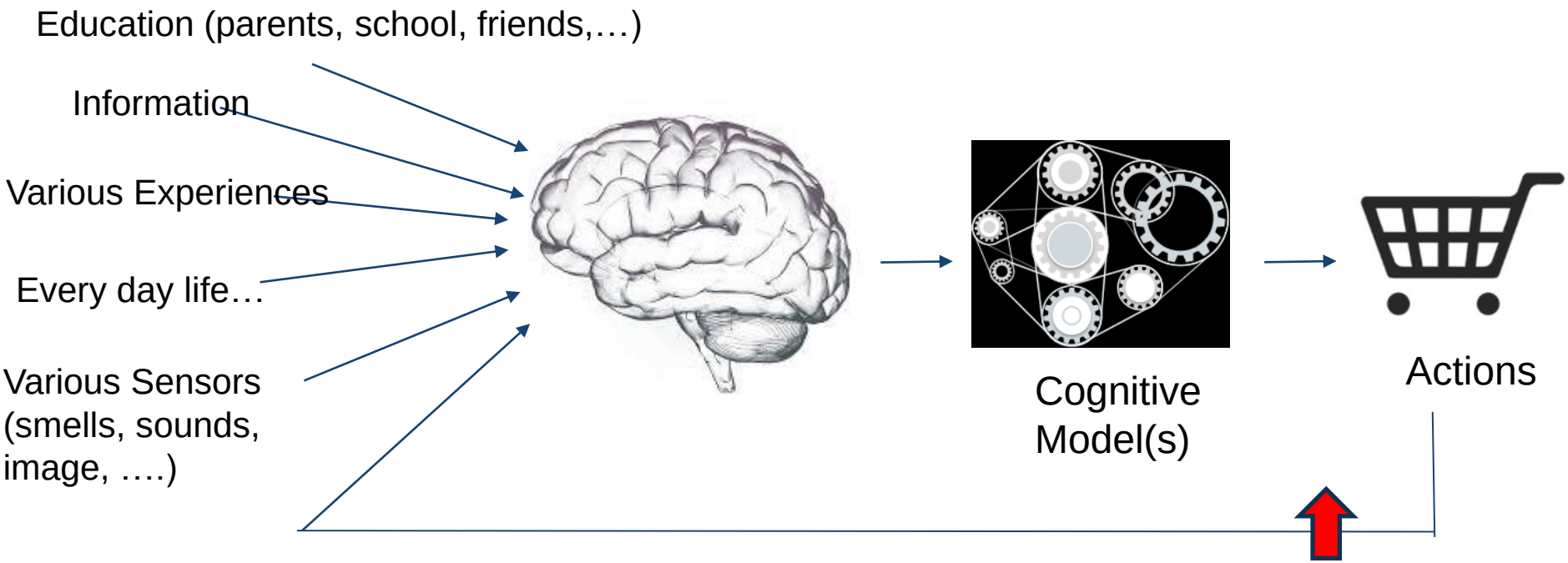
Cambridge Analytica on Steroids with chatGPT

C

write a short advertising message to convince an intravert neurotic person to buy a pair of Nike shoes

"Elevate your comfort and confidence quietly with Nike. Our shoes are designed for your unique journey. Step into serenity and conquer the world on your terms. Order your Nike shoes now, and discover the power of inner strength."

Action Execution Attacks



(4) Execution Attacks:
*Attacks that **prevent** the subject to execute his actions (**Dark Patterns + social engineering**)*

Dark Patterns

Making a process more difficult than it needs to be, with the intent of dissuading certain action(s).

Dark Patterns

Politique cookies

Continuer sans accepter →

Notre organisation et ses partenaires stockent et/ou accèdent à des informations sur votre appareil, telles que les identifiants uniques de cookies pour traiter les données personnelles. Nous procédons ainsi pour fournir un contenu personnalisé. Vous pouvez accepter ou gérer vos préférences en cliquant sur Afficher les finalités ou à tout moment sur la page de la politique de confidentialité. Vos préférences renseignées sur ce site sont signalées à nos partenaires. **Politique Cookies**
Nous utilisons des cookies et technologies similaires tiers ou non pour prévenir les risques de fraude.

Utiliser des données de géolocalisation précises. Analyser activement les caractéristiques du terminal pour l'identification. Stocker et/ou accéder à des informations sur un terminal. Publicités et contenu personnalisés, mesure de performance des publicités et du contenu, données d'audience et développement de produit.

Liste de nos partenaires

GÉRER MES COOKIES | ACCEPTER

Social Engineering and Generative AI

AI-enabled fraud will fundamentally change the attack surface in following:

- Crafting more convincing phishing emails and messages that closely mimic legitimate sources (phishing)
- Deepfakes focusing mainly on voice cloning (Vishing)
- AI-driven data mining: using AI to identify potential targets and to determine the most effective approach for a social engineering attack, thus increasing the likelihood of success

(see ENISA Threat Landscape 2023 report)

Artificial Intelligence and Manipulation

- Information manipulation is not new…Automation and the use of AI is!
- This automation allows to launch large and personalized manipulation campaigns that are very efficient and at a low cost!
- With AI, manipulation will improve… but AI can also (maybe) help detection these manipulations!
- We need more research and strong regulations:
  - DSA (Digital Service Act) will regulate “Dark Patterns”
  - IA Act will regulate AI (in particular emotion detection, manipulations,…)

Merci!

claude.castelluccia@inria.fr