Panel 1 - Human-machine interaction and human control

Geber RAMALHO

Background

- Electronic engineer (1988)
- PhD in Artificial Intelligence – Paris VI (1997)

Currently positions

- Professor in Computer Science Center - UFPE
- Chairman of the board of CESAR Institute

Interests

- AI for art and entertainment
- Ethics and AI
- Innovation and entrepreneurship

What would be an ethical AI?

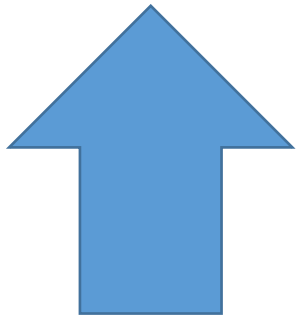
How to guarantee that a given intelligent system will have an ethical behavior?

Luciano Floridi's Principles

1. **Beneficence**: promoting well-being, preserving dignity and sustaining the planet
2. **Non-maleficence**: privacy, risk and misuse prevention, "capability caution"
3. **Autonomy**: the power (of the user) to decide (or not)
4. **Justice:** promoting prosperity and preserving solidarity
5. **Explicability** (giving machine decisions intelligibility and responsibility

Ban LAWS!

- **In some cases?**
- **Under certain circumstances?**
- **For some weapons?**



How to “limit the damage”?

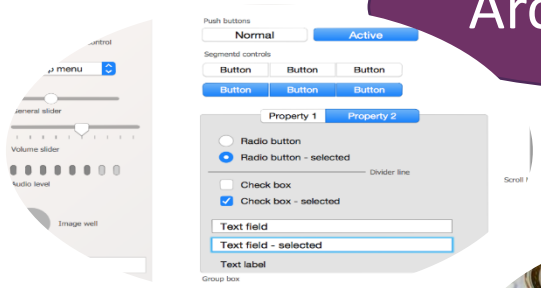
Which are the adoption critera, processes, responsibilities?

Some insights from our research group on Ethics and AI at UFPE

Regulation and the pathetic dot framework
(Lawrence Lessig, 1999)

Market

Architecture



Norms

Law

ISO

Law: Criteria for the adoption of fully automated AI

- Preliminar work
- Identify criteria for adopting HOOTL (Human out of the loop) approach in 3 (regulated or requiring regulation) domains
  - Intensive Care Unities
  - Electricity distribution
  - Lethal Automated Weapons
- Compare them looking for convergence

Intensive Care Unities

Case	Description	Examples	Interaction
Resuscitation	Immediate intervention to save life	- Cardiac arrest - Massive bleeding	HITL
Emergency	High risk of deterioration (leading to death) or signs of critical problems	- Chest pain (cardiac) - Asthma Attack	HITL
Urgent	Stable but requires multiple resources for diagnosis and treatment (laboratory tests, X-rays, tomography, etc.).	- Abdominal pain - High fever with cough	HOTL
Slightly urgent	Stable requiring few resources (a simple X-ray or sutures).	- Simple laceration - Pain when urinating	HOTL
Not urgent	Stable without need for resources beyond the prescription	- Abrasion - Renew medicine	HOOTL

LAWS

- Target precision (distance) α HOOTL
- Reponsibility/Explicability α HOOTL
- Damage Extent 1/α HOOTL
- Context/Environment complexity 1/α HOOTL
- Dignity (human as target) 1/α HOOTL

Comparison: criteria influence for adopting HOOTL

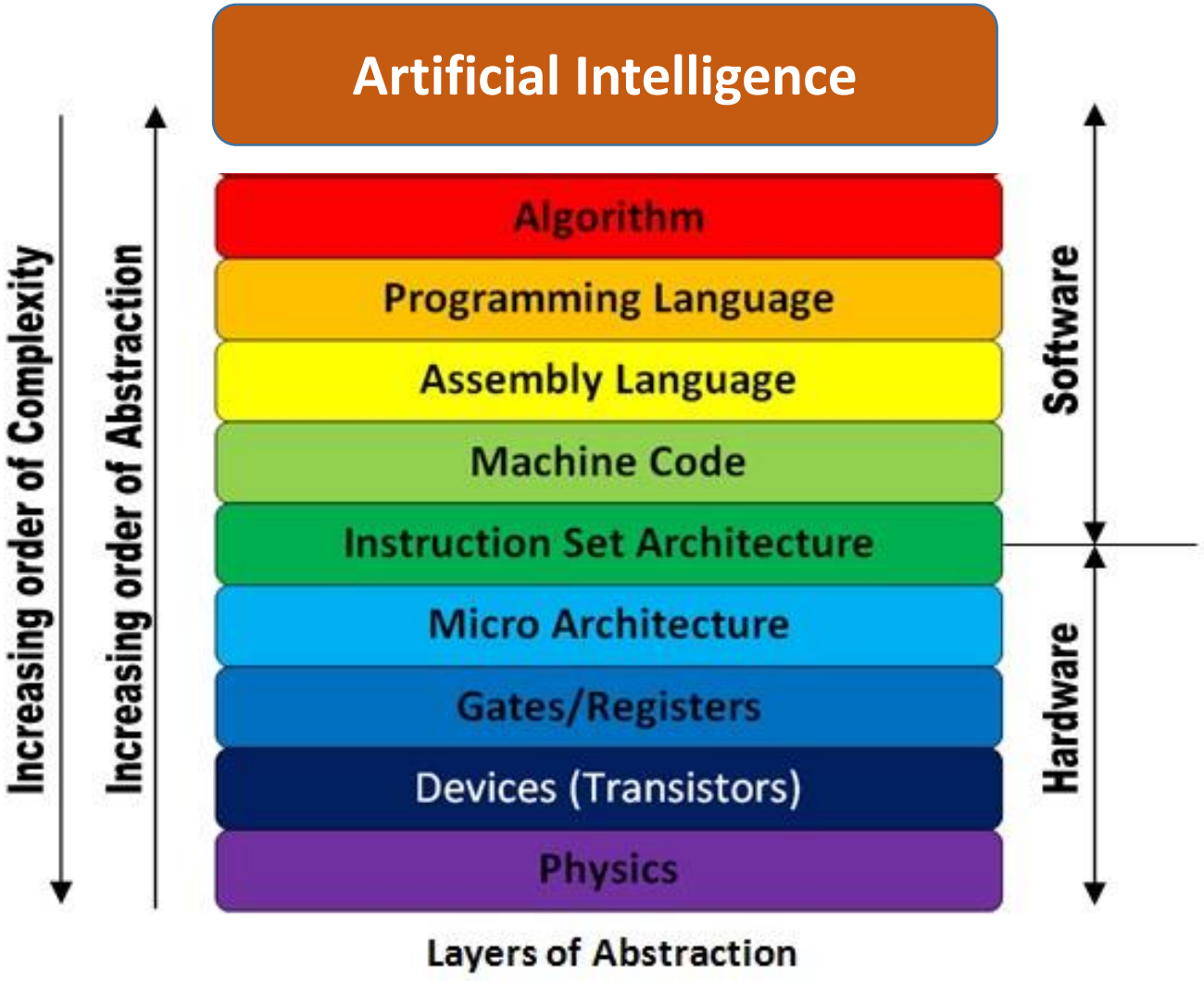
Time to act	α	1/α	1/α
Impact on people	1/α	α	1/α
Cost		α (operation)	α (troop life)
Responsibility	1/α	1/α	1/α

Market: certifications

- Ethical AI for enterprises (similar to the B-system and “great place to work”)
  - CRISP-DM process vs. Floridi’s principles => 48-questions questionaire

	Business understanding	Data understanding	Data acquisition	Data preparation	Modeling	Evaluation	Deployment	Observation in the wild
Beneficence: Promoting Well-Being, Preserving Dignity, and Sustaining the Planet	Ethical business goals + impacts	Guarantee that all populations are represented equally in the data	unpleasant data request + compliance	unpleasant data exclusion, stratified samples, data balancing	Transfer learning; Escolher modelos energeticamente eficientes, data parameterization	Assess whether the model has deviated from initial beneficent goals during the development process		
Non-maleficence: Privacy, Security and "Capability Caution"	Risk planning; System legal compliance	using data with potential misuse	Using data only when under affirmative consent; Deleting data and erasing traces when consent is revoked; Not buying personal user data	Data protection through ISO 27000 serie	Design rigorous testing processes for applications that deal with sensitive data	Impacts evaluation, regression analysis		
Autonomy: The Power to Decide (Whether to Decide)	Automation impact assessment; Critical areas / decisions;	social discrimant propagation	Automatic dataset increment	Probing correlations in data, removing sensitive data and their proxies	unbalanced errors considerations, local minimium use, isolated examples exclusion	Need for model efficiency monitoring		
Justice: Promoting Prosperity and Preserving Solidarity	Doesn't allow for poverty + social entrapment;	Legal compliance	unbalanced data acquision	Pertinent demographic groups are represented in equal proportions on the training and testing datasets	Design tests aimed at detecting unfair treatment of demographic groups, analyze the robustness	Assess whether the model discriminates against demographic groups	Não contrariar restrições locais.	
Explicability: Enabling the Other Principles Through Intelligibility and Accountability	Business model visibility; Business model updated.	Data correlation and importance	Data lake acquisions	Data preprocessing record	Surrogate models;	Document mining process in final report; confusion matrix analysis	Ethical requirements (RNF) must be materialized and implemented.	Ethical Commite

Architecture: abstract layers in computing



Rules + reasoning > goals > learning

```
60 PRINT S $
70 INPUT "Do you want more stars?"; Q $
80 IF LEN (Q $) = 0 THEN GOTO 70
```

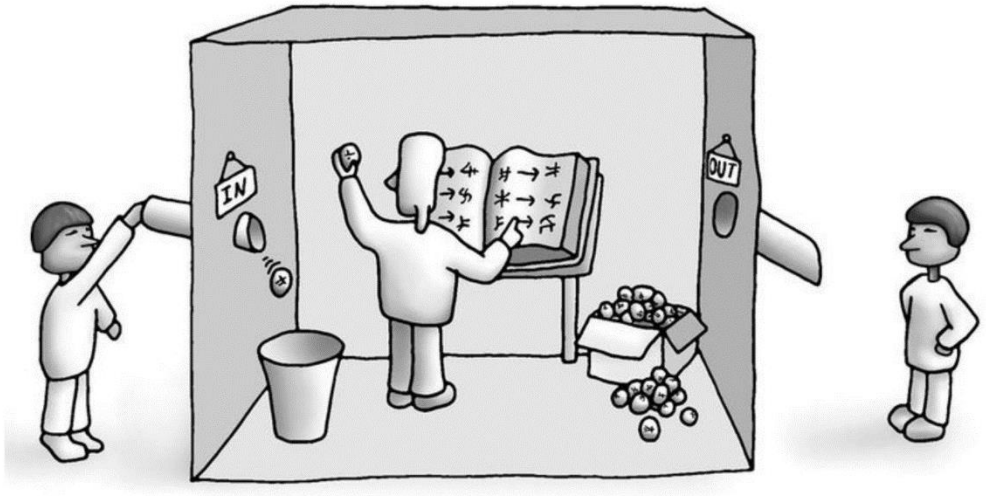
```
//J = 25;
MOV R3, #25
STR R3, [R11, #-12]
```

```
F0 FE 14 04 1C 70 04 A0 D0 80 EF 00 70
FB C0 50 D8 F7 00 00 BB EF E0 F0 D1 00
```

The more abstract, the easier to program, but less control you have!

AI limitations: reasoning

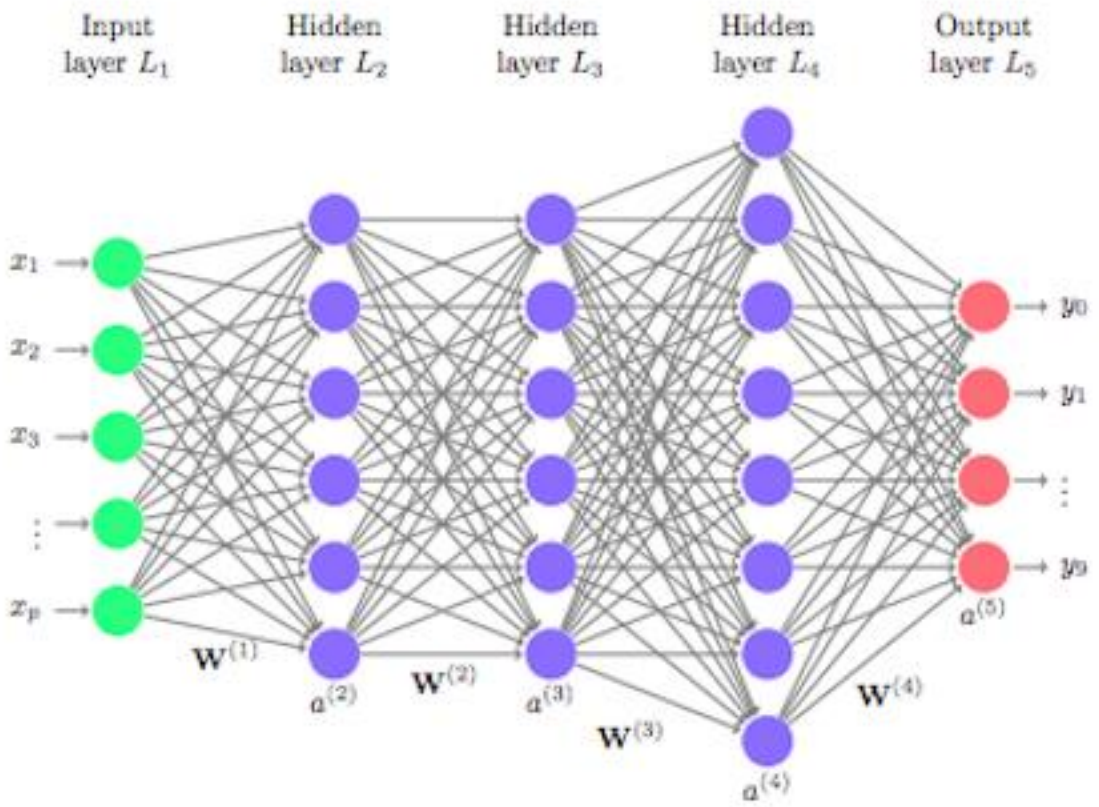
A) $\forall$ x,y,z Americano(x) $\wedge$ Arma(y) $\wedge$ Nação(z) $\wedge$ Hostil(z) $\wedge$ Vende(x,z,y)
$\Rightarrow$ Criminoso(x)
B) $\forall$ x Guerra(x,USA) $\Rightarrow$ Hostil(x)
C) $\forall$ x InimigoPolítico(x,USA) $\Rightarrow$ Hostil(x)
D) $\forall$ x Míssil(x) $\Rightarrow$ Arma(x)
E) $\forall$ x Bomba(x) $\Rightarrow$ Arma(x)
F) Nação(Cuba)
G) Nação(USA)
H) InimigoPolítico(Cuba,USA)
I) InimigoPolítico(Irã,USA)
J) Americano(West)
K) $\exists$ x Possui(Cuba,x) $\wedge$ Míssil(x)
L) $\forall$ x Possui(Cuba,x) $\wedge$ Míssil(x) $\Rightarrow$ Vende(West, Cuba,x)



M) Possui(Cuba,M1) - *Elimination of existential quantifier and the conjuction in K*
N) Míssil(M1) - instantiation
O) Arma(M1) - *Modus Ponens from D e N*
P) Hostil(Cuba) - *Modus Ponens from C e H*
Q) Vende(West,Cuba,M1) - *Modus Ponens from L, M e N*
R) Criminoso(West) - *Modus Ponens from A, J, O, F, P e Q*

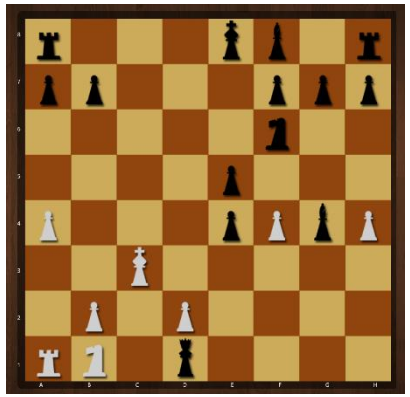
AI limitations: explicability

- Sometimes decisions cannot be explained!



AI limitations: one task-oriented

- AI has good performance in narrow application domains



The story of Carlsen winning the "double," getting the triple crown and finishing the year as the world champion and world number-one in standard, rapid and blitz is big. However, the incident on the last day in his game with **Alireza Firouzja**, who lost on time and whose protest was rejected, boosted the comments even further, and this story just makes it into the top-10! **208 comments** (at the time of writing!).

Architecture: AI limitations must be tracked and stated clearly

GGE principle (g) “Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems”

How to translate this into a practical measure?

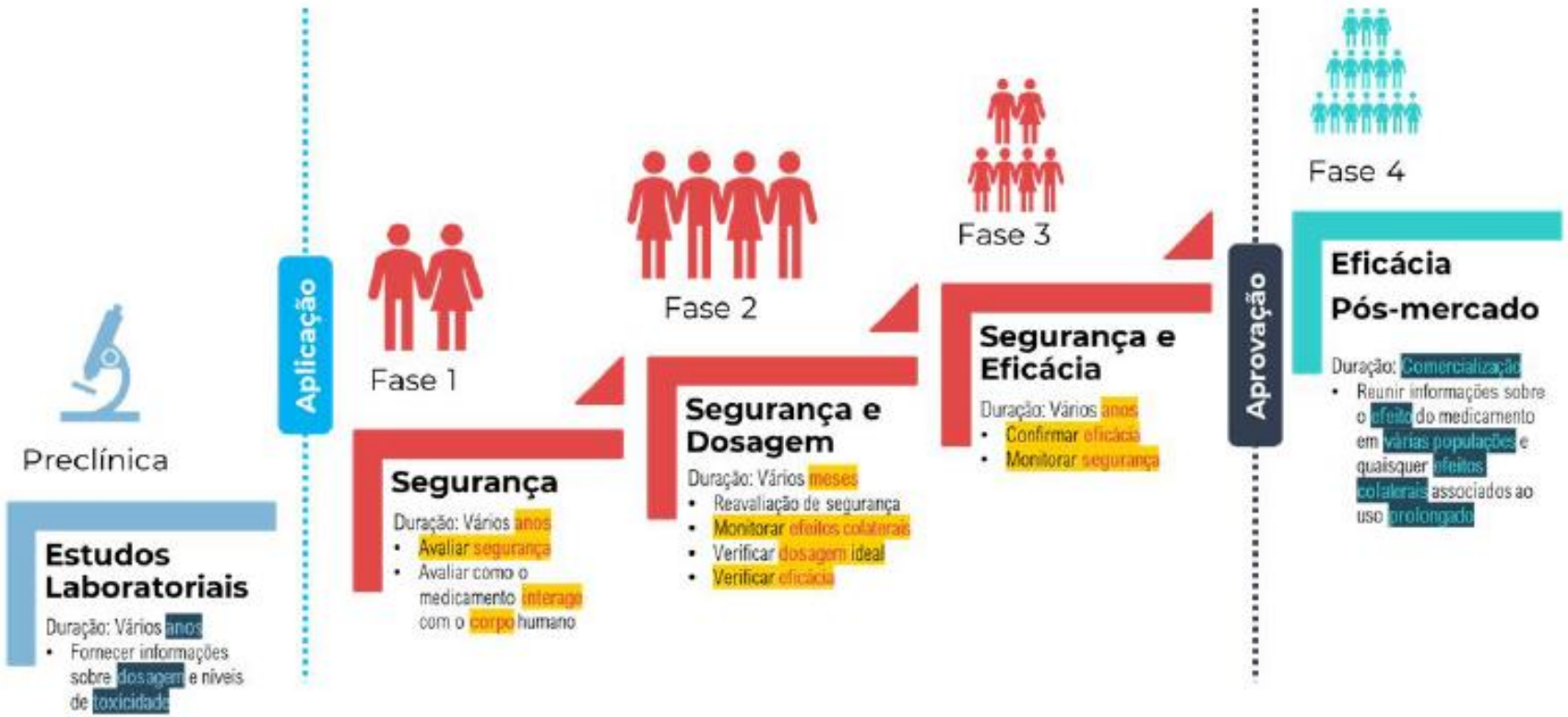
Consumer Artificial Intelligence Information (CAII)

- A parallel with pharmaceutical industry!
  - FDA (US), ANVISA (Brazil), TGA (Australia)
- Concerning drugs
  - It is approved through a long process of tests
  - We know a lot of things (18 items): eficacy, side-effects, constraints, dosage...
  - The Prescribing Information is a contract

Resumo
Indicações e Uso
Dosagem e administração
Formas e dosagens de dosagem
Contraindicações
Avisos e Precauções
Reações adversas
Interações medicamentosas
Uso em populações específicas
Abuso e dependência de drogas
Sobredosagem
Descrição
Farmacologia Clínica
Toxicologia Não Clínica
Estudos clínicos
Referências
Como fornecido / armazenamento e manuseio
Informações de aconselhamento ao paciente

Consumer Artificial Intelligence Information (CAII)

- The research, approval and deployment process for AI systems



Technology is part of the problem, but may be part of the solution!

Algorithms for mitigation of bias

- Optimized Preprocessing (Calmon et al., 2017)
- Disparate Impact Remover (Feldman et al., 2015)
- Equalized Odds Postprocessing (Hardt et al., 2016)
- Reweighing (Kamiran and Calders, 2012)
- Reject Option Classification (Kamiran et al., 2012)
- Prejudice Remover Regularizer (Kamishima et al., 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., 2017)
- Learning Fair Representations (Zemel et al., 2013)
- Adversarial Debiasing (Zhang et al., 2018)
- Meta-Algorithm for Fair Classification (Celis et al.. 2018)

Consumer Artificial Intelligence Information (CAII)

- 41 information items covered
- About AI
  - Intended use
  - Explanations
  - Model resource
  - Algorithms
  - Training data
  - Training environment
  - Optimizartion goals
  - User intarface
- About data
  - Sensors and sources
  - Actuators and outputs
- Legal Information
  - Lead programmer
  - Registration
  - Developed by
  - Consumer contact
  - Impact report
  - …

Consumer Artificial Intelligence Information (CAII)

- Consumer information
  - What should I know to use the system?
  - How my data will be used?
  - Where and for how long my data will be stored?
  - Who my data will be shared with?
  - When my data will be shared?
  - ...

Main messages

- Fully automated weapons may perahps be inevitable, but risks should be cotrolled and technology + regulation can help
- Law is not the only possible regulation, and sometimes not the best one
- It is worth looking at what is being discussed in ethics and AI in general

Final remarks on GGE's principles

- (b) **Human responsibility** for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines.
- **Who? Define clearly the stakeholders!**