Beneficial AI: the next battlefield

Eugénio Oliveira
eco@fe.up.pt

"In an A.I.-first world, we are rethinking all our products," Sundar Pichai said."
The New York Times, May 18, 2017

INTERNATIONAL CONFERENCE ON
ARTIFICIAL INTELLIGENCE AND INFORMATION
December 6, 2017

OUTLINE

- AI definition, life-line and Achievements
- Hyping AI
- Artificial General Intelligence
- Emotions and Sentiments ?
- Simplistic remedies
- ART in Artificial Intelligence ! And recommendations …

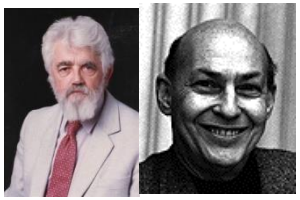
Beneficial AI, the next battlefield

"after many **small steps** for AI researchers, will it result in a **giant leap** in the unknown for mankind?"

AI Life-line

"We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire.

The "motto"


J.Mc.. M.M.

*The study is to proceed on the basis of the **conjecture** that every aspect of learning or any other feature of **intelligence can** in principle be so precisely **described** that a machine can be made to **simulate** it. An attempt will be made to find how to make **machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves**.*

*- Dartmouth AI Project Proposal; **J. McCarthy** et al.; Aug. 31, 19**55**."*

Defining AI

- ***Intelligence*** can have many faces like ***creativity***, ***solving problems***, ***pattern recognition***, ***classification***, ***learning***, ***induction***, ***deduction***, ***building analogies***, ***optimization***, ***surviving in an environment***, ***language processing***, *knowledge and many more. A formal definition incorporating every aspect of intelligence, however, seems difficult.*

"Most, if not all known facets of intelligence can be formulated as goal driven or, more generally, as maximizing some utility function.

Marcus Hutter, Universal Artificial Intelligence, Springer

If true, the problem would be: Which Function?

Defining AI

*"theory and development of computer systems able to **perform tasks** normally requiring human intelligence"*

*"a branch of computer science dealing with the **simulation of intelligent behaviour** in computers"*

*"the capability of a machine to **imitate intelligent human** behaviour"*

ML Five Tribes

- **Symbolists**
- **Connectionists**
- **Evolutionaries**
- **Statisticians (Bayesians)**
- **Analogizers**

Different schools, same objectives:
To develop machine intelligence!

The Master Algorithm: The ultimate Learning Machine that will remake our world

(P.Domingos, Basic Books, 2015)

AI achievements

Symbolists in the 80ies and 90ies **KBS**

- Deductive reasoning
- Knowledge representation
- Uncertainty

Theoretical topics like:

- Computational Logic;
- Fuzzy, Non-Monotonic, Intentional, Modal Logics

AI achievements

- In 2006 R. Brooks opposed the idea that AI had failed and warned that AI would be **around us** every day.

(in Stefan Wess, "AI: dream or nightmare ", TEDx Zurich Talks, 2014)

AI programs **predict** everything from our taste in music to our likelihood of committing a crime, to our fitness for a job or an educational opportunity.

Current AI-based systems allow:

- Facebook to decide which updates to show to the user and Twitter to decide which twits to show
- Companies to record clients profile
- Satellites to be intelligently managed

AI achievements

- **Deep Blue and AlphaGo**

(In October 2015, AlphaGo became the first Computer Go program to beat a professional human Go player without handicaps on a full-sized 19×19 board.)

- **Google** does page ranking for us
- Computers perform Intelligent spam **filtering**
- Amazon and Netflix **Recommender Systems** are in use
- PANDORA **records** our preferred kind of **music** from radio
- INRIX helps to select the **best route** during traffic jams
- BingTravel **predicts** future **flights price**
- BDI **Agents** representing us in the **stock exchange**
- Skin steins can be automatically **recognized** as carcinogenic

AI achievements

- NSA's algorithms decide whether you're a potential terrorist.
- Climate models decide what's a safe level of carbon dioxide in the atmosphere.

P.Domingos, The Master Algorithm

Toby Walsh on "How can you stop killer robots" at TEDxBerlin (http://www.tedxberlin.de)

- *"The new wave of A.I.- enhanced **assistants** — Apple's Siri, Facebook's M, Amazon's Echo — are all creatures of machine learning, built with similar intentions. … **the greatest danger** is that the information we're feeding them is biased in the first place."* *Sundar Pichai, CEO of Google*

Machine Learning (ML) algorithms work together with a multitude of other algorithms in order to get the things done

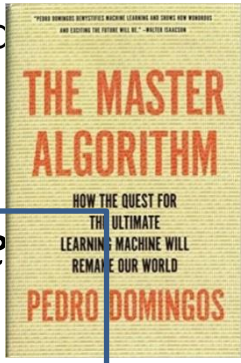
- ***Siemens Healthineers and IBM Watson Health*** *are tackling population health together. By combining the clinical expertise of Siemens and cognitive computing leadership of IBM Watson Health, you can make critical healthcare data meaningful and create actionable initiatives that will help your organization thrive.*

Good

- *"Google's* ***self-driving car*** *taught itself how to stay on the road; no engineer wrote an algorithm instructing it, step-by-step, how to get from A to B"*

Problematic?

- 300 of a patient words enable **Watson** the detection of a future (some kind o psychosis



Hyping AI

- *If it exists, the **Master Algorithm** can derive all knowledge in the world—past, present, and future—from data. Inventing it would be one of the greatest advances in the history of science. P.D., TMA*

- The **Master Algorithm** as a combination of **current** ML algorithms working over big data → "**ultimate learning machine**".

Overselling

- through the "**Master Algorithm**", by mining a "*vast amount of patient data and drug data combined with knowledge mined from biomedical literature is how we will **cure the cancer***".

Overselling or Optimism?

- "Big data is not the new oil; it's the new snake oil".

Unfair Reaction?

Hyping AI

Super Inteligence? Artificial **General Intelligence**? **Strong AI**?

Distraction?

Last years rupture:



GH / YLeC/ANg

"Deep Learning" + "Big Data"+ Network computing power

AI current CLAIMS

- *"**Machine learning** is remaking science, technology, business, politics, and war...."* P.Domingos, TMA

- *" High resolution brain scanners will sense from electro-magnetic fields of each one of your neurons and the world around you will adapt continuously in response, and thanks to ML which lets computers predict the future based on past experience the **world will guess what you want before you even want it and will be ready for you"*** (P.Domingos TEDxLA 2016).

Scaring?

Hyping AI

- in 40 or 50 years …*"Boundaries between self and others will begin to dissolve."*
- *"The Question what means to be human will no longer have an answer. But may be it never did."*

In P.D. "Next 100 years of your life" TEDxLA Talks 2016

- We are not here talking about claims by either a scientific fiction writer or even a philosopher, but of a CS professor at an USA prestigious university.

"AI systems are already making **problematic judgements** that are producing significant social, cultural, and economic impacts in people's everyday lives."

Why we urgently need to measure AI's societal impacts
By Kate Crawford and Meredith Whittaker

it is an **iconic portrait** of the indiscriminate horror

Danger!
TAY a Twitter's Chatbot by Microsoft : learnt that "Hitler did nothing wrong"

Artificial General Intelligence

Deep Blue and Alfa Go: intelligence without consciousness?

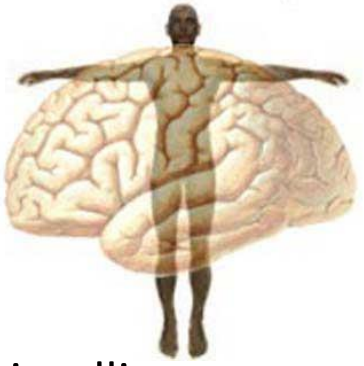
Turing Test? It depends …

John Searl and the “Chinese Room” argument

Argument OR Paradox?

“What is indeed a movie?” Jean Tardy, author of “The Meca Sapiens Blueprint “
“The Creation of a Conscious Machine”

Artificial General Intelligence raises the problem of a possible artificial Consciousness



Dualism versus Monism
Cartesians are Dualists
Science current trend is basically Monistic (Christian von Wolff)

Will Consciousness be an EMERGENT property of many specific intelligences at work??

I do not know. A real Digital Mind in the **near** future it is not probable.

But "Strong AI" can exist and raise many ethic problems anyway

Does the mind work like a computer? What about the opposite?

The mind *usually* resides on the Brain (Hardware)

Intelligence and Autonomy are still distinctive for the Human and the Artificial

Different infrastructure:

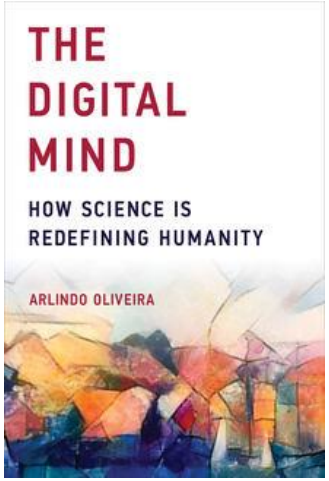
Parallelism Versus Sequencing (like for logic inferences)

But Similar basic capabilities:

Decision Making (reasoning, fuzzy, by default,..) Interaction(text, speech)

Submarines do not swim; planes do not flap wings …

The Mind is an emergent property of the Brain enabling humans to have a set of cognitive capabilities including intelligence, consciousness, free willing, reasoning, memory, emotions, etc.
"The Digital Mind", Arlindo Oliveira



Does the mind work like a computer? What about the opposite?

To mimic a mind "in silico" Versus "in vivo".

Global architecture of the Brain plays a role in Intelligence. Brain modules are the nodes of a conection network including multiple structures.
"connections change dynamically" and " **inteligence reflects the capability to be flexible and easily go from one network to the other"** *Trends in Cognitive Sciences, Nov. 2017*
"Network Neuroscience Theory of Human Intelligence", Aron Barbey, University of Illinois

Watson won "Jeopardy "→factoids. BUT…
Was it happy after winning?

Sentiments are different (but come after) **Emotions** (AD says)
We are already able to represent kind of basic Emotional states (fear, anxiety, …) in Software Agents.
Those Emotions can change the way Agents reason, memorize, act, decide, for a certain period of time while that emotional state lasts.
We may say the agents act differently when they "feel like" happy or fearful or …

ℰBDI LOGIC

The basic Emotional-BDI Logic system is characterised by the union of the following **set of axioms**:

1. the set of all propositional **tautologies**
2. the **time** axiom set CTL
3. the **action** axiom set PDL
4. the **belief** axiom set BEL_{KD45}
5. the **desire** axiom set DES_{KD}
6. the **intention** axiom set INT_{KD}
7. the **capabilities** axiom set CAP
8. the **resources** axiom set RES
9. the **fear** axiom set $FEAR_{K}$
10. the **fundamental desire** axiom set $FDES_{KDT}$

we define an **ℰBDI-model** with a set of modal operators:
Op = {BEL,DES, INT, FDES, FEAR, …},

-David Pereira, Eugénio Oliveira and Nelma Moreira: "***Formal Modelling of Emotions in BDI Agents***" *published in Computational Logic in Multi-Agent Systems*, Lecture Notes of Computer Science, V.5056, 2008

- "***Modelling Emotional BDI Agents***" *David Pereira, Eugénio Oliveira, and Nelma Moreira*, in Formal Approaches to Multi-Agent Systems, Ed. Barbara Dunin-Keplicz and Rineke Verbrugge, ECCAI Workshop Proceedings, pp.47-62., Riva del Garda, Italy.

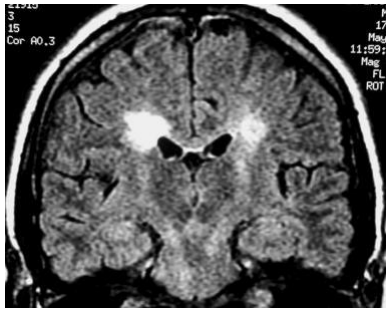
Threats represent facts or events occurring in the environment which directly affect one or more **fundamental desires** of the agent, putting at stake its self preservation
Danger: a threat is dangerous when the agent believes that some condition leads inevitably to the falsity of a **fundamental desire** φ, and also believes that it will also be inevitably true in the future.

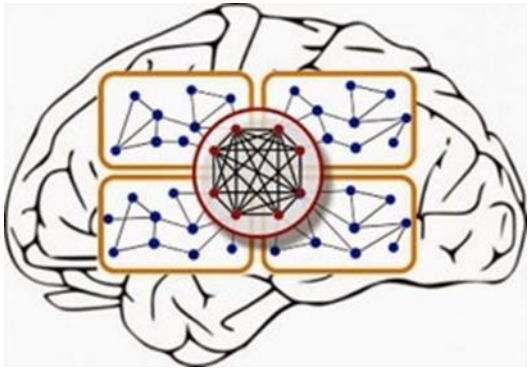
$$DangerousThreat(\Psi,\varphi) \equiv FDES(\varphi) \wedge BEL(\Psi \rightarrow AF(\neg\varphi)) \wedge BEL(AF\ \Psi)$$

$$StronglyUnpleasant(\Psi,\varphi) \equiv DES(\varphi) \wedge BEL(\Psi \rightarrow AF(\neg\varphi)) \wedge BEL(EF\ \Psi)$$

Can a computer work like a Mind?

Download a mind to an artificial entity (robot, computer, computer network)?
Reverse Engineering… ?





MRI, …, too much superficial.
Resolution of $1mm^3$… 50-100K neurons, 10^2 to 10^3 synapses…
We need interpretation at the Nano metric level

Evolution of Digital Brain →very complex stimula → huge number of diverse sensors .

The right answer is not yet available !!!!

Human-like AI?

How far away are we from building "human-like" AI? What are the key problems that we need to solve before we get there? George Zarkadakis, Digital Lead at Willis Towers Watson, LinkedIn,10/22/2017

Human-level AI: The roadmap

Need to Solve for	State of Play
Generality	Solved ✔
Learning without being taught	olved ✔
Transfer learning	Not yet
Common sense	Not yet
Self-awareness	Still mysteriou

really a **hard** problem. The function of human **memory** is perhaps the key to developing **common sense** in machines

DL networks (as opposed to "symbolic", and ES), have demonstrated **generality**

RL used with AlphaGo by DeepMind demonstrated how a ANN that is given a goal can learn and **invent** strategies

use, or **abstract**, the knowledge accumulated by solving a specific problem, and apply this knowledge in solving a **different** problem.

Machines that will have **us believe** they **have a self, or a personality**, should be relatively easy to develop. But whether they would be truly self-aware, we will only know if we crack the "hard problem of consciousness" first.

- …things might **go wrong!** There's no sensationalism here, this is a … and pragmatic discussion.
- ***reinforcement learning*** **(RL),** in which agents … environment:

One counter-measure is to penalise "changes to the environment"

  - for an **agent** operating in a large, multifaceted **environment**, an **objective** that focuses on only one aspect of the environment may implicitly express **indifference** over other aspects of the environment. An agent optimising this objective function might thus engage in major **disruptions** of the broader environment if doing so provides even a tiny advantage for the task at hand.

    - **exploration** can be **dangerous**, since it involves taking actions whose consequences the agent doesn't understand well.

- the increasing trend towards fully autonomous systems points towards the need for a **unified approach to prevent** these systems from causing unintended harm.

"Concrete problems in AI safety", Amodei, Olah, et al., arXiv 2016

What can we do? **Simplistic** remedies

Possible **naïve** answers are:

- **Just remove the plug!**

Where is the plug (in the cloud)?

- ***Always have a kill switch*!**

Intelligent machines will learn to preserve themselves.

- **Put the machine into a "cage".**

That is what a Virtual Machine is. Not safe enough even to prevent many smart virus.

- Nature deals with exponential growth of bacteria, virus, animals and diseases **creating natural enemies** to fight.

Multiplying the problem?

Based on Stefan Wess, "AI: dream or nightmare" TEDx Zurich 2014

What can we do? **Simplistic** remedies

- May be the way to deal with the problem is **not** about technology.
- May be it should be about **Machine ethics**.
- How to make intelligent machines sharing our values?

- **Legislation** applied to people responsible for AI systems design, dissemination and application

- Elon Musk at the World Government Summit:

[*we may become irrelevant as artificial intelligence (AI) grows more prominent*], February 2017, Dubai

What can we do? **Simplistic** remedies

PAST EXPERIENCE

Dealing with critical situations (using MAS for D-Making)**:**

- to manage and coordinate actions:
  - on a ship under attack
  - When unexpected disruptions happen in an Airline OCC

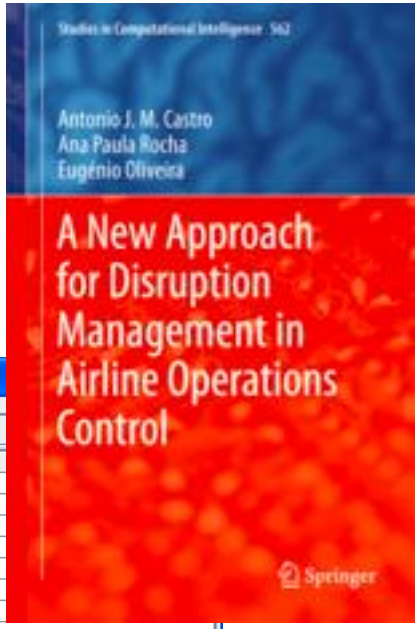
it should be mandatory that the automated system **specification** enforces the **human in/on the loop**. Someone who will be **accountable** for the most important decisions in run-time.

PAST EXPERIENCE

- How do humans react to **dangerous** tasks like firefighting?
- Human-like agents (Softbots / Robots) → **Emotion**-based
- **Emotional** states result from human **evolution** not as a kind of reasoning by-pass but as complementary machinery to make **humans to better deal with critical situations**.
- Why not include a cautious **emotion-like** mechanism in the **artificial agents**?
- Computational **Trust** Models to select the right partners in a Netw.

PAST EXPERIENCE

Ex: MASDIMA for Airline Operations Disruption Management



MASDIMA - Multi-Agent System for Disruption Management

Airb...	Flight	Origin	Departure	Destin...	Arrival
	733	BCN	2009-09-02 13:35 U...	OPO	2009-09-02 15:2...
	851	FCO	2009-09-02 13:40 U...	OPO	2009-09-02 16:4...
	926	LIS	2009-09-02 13:40 U...	ZRH	2009-09-02 16:3...
	228	LIS	2009-09-02 13:50 U...	CMN	2009-09-02 15:2...
	1970	LIS	2009-09-02 13:50 U...	OPO	2009-09-02 14:4...
	804	LIS	2009-09-02 13:50 U...	MXP	2009-09-02 16:3...
	614	LIS	2009-09-02 14:00 U...	BRU	2009-09-02 16:4...
	193	LIS	2009-09-02 14:00 U...	GRU	2009-09-03 00:2...
	165	LIS	2009-09-02 14:00 U...	FOR	2009-09-02 21:3...
	747	BCN	2009-09-02 14:00 U...	LIS	2009-09-02 16:0...
	1633	LIS	2009-09-02 14:15 U...	FNC	2009-09-02 16:0...
	422	LIS	2009-09-02 14:20 U...	TLS	2009-09-02 16:2...
	484	LIS	2009-09-02 14:20 U...	NCE	2009-09-02 16:4...
	440	LIS	2009-09-02 14:20 U...	ORY	2009-09-02 16:5...
	944	LIS	2009-09-02 14:20 U...	GVA	2009-09-02 16:5...
	410	LIS	2009-09-02 14:40 U...	MRS	2009-09-02 17:0...
	675	LUX	2009-09-02 14:45 U...	OPO	2009-09-02 17:1...

Aircraft | Crew | Pax and Airport

Attribute	Value
Tail Number	CSTMW
Name	AIRBUS A320-211
Model	A320
Fleet	NB
Maximum Take-Off Weight	73500
Maintenance Average Cost per Minute	10.330
Air Traffic Control Cost per Nautical Miles	2.880
Fuel Average Cost per Minute	16.520
Airport Handling Average Cost per Day	950

Flight Affe...	Estimated Departure	Delay	Violations	Status
944	2009-09-02 14:35 UTC	15.0	9	Solved
856	2009-09-04 07:28 UTC	23.0	8	Solved
608	2009-09-04 09:30 UTC	30.0	10	Solved
1970	2009-09-04 14:35 UTC	45.0	18	Solved
1917	2009-09-04 22:00 UTC	35.0	8	Solved
712	2009-09-08 08:22 UTC	17.0	8	Solved
614	2009-09-08 14:15 UTC	15.0	8	Solved
834	2009-09-09 06:25 UTC	10.0	9	Unsolved

Solution Plan | Supervisor Default Values

Violations | Solution Proposal

Attribute	Value
Aircraft Delay	7.0
Aircraft Cost	420.0
Crew Delay	49.0
Crew Cost	227.0
Passenger Delay	7.0
Passenger Cost	26.0
Solution Utility	0.867 (86.7%)

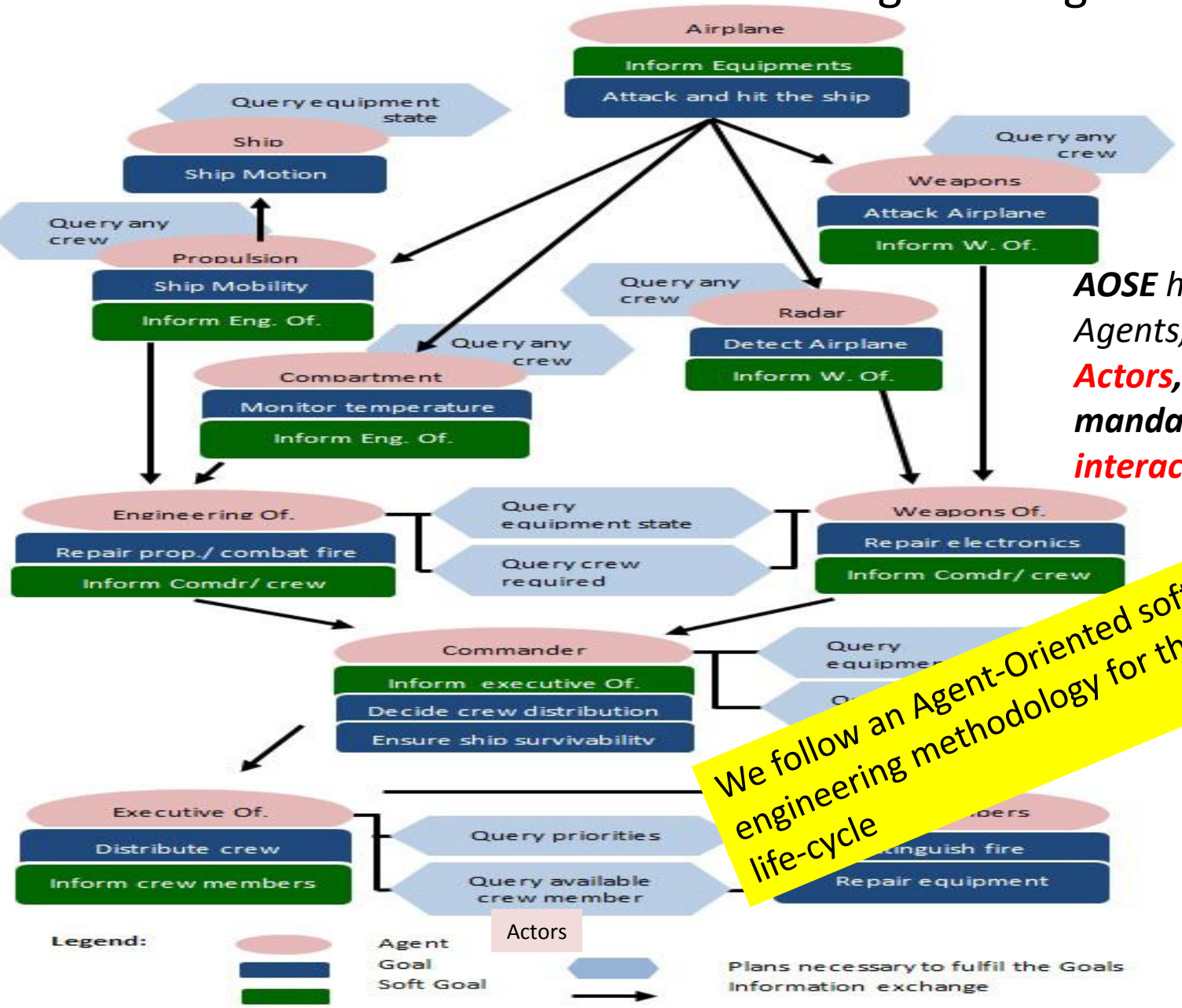
PAST EXPERIENCE

Ships subjected to external factors, such as environmental conditions, external attacks and potential damages, perform several needed tasks defined by its operational requirements

(Crew assignment to tasks; Tasks prioritization

A **Multi-Agent System** to automatically manage these situations has been specified such that the chain of command is replicated and **command officers legitimate** some crucial **decisions**

Actors and goals diagram



AOSE *helps on specifying Agents, hard/soft Goal,* ***Actors, Roles*** *and* ***mandatory human interaction***

We follow an Agent-Oriented software engineering methodology for tha AI system life-cycle

AI Now Institute at New York University organized AI Now 2017 Symposium and Workshop

- **AI Now 2017 Report**

Ethical questions of **bias and accountability** will become even more urgent in the context of rights and liberties as AI systems capable of **violent force** against humans are developed and deployed in law enforcement and **military contexts**.

Peter Asaro has pointed to difficult ethical issues involving how **lethal autonomous weapons systems (LAWS)** will detect threats or gestures of cooperation, especially involving vulnerable populations. He concludes that **AI and robotics researchers should adopt ethical and legal standards** that maintain human control and accountability over these systems.

What can we do?

AI NOW 2017 Report

The AI NOW 2017 Symposium reflecting on policy interventions deeply examined the near-term **social and economic implications of AI** addressing the following key issues:

- **Labour Automation**: AI and automation impact on labour practices (employee hiring, firing and management);
- **Bias and inclusion**: Bias that AI systems may **perpetuate** and even amplify due to **biased training data** and faulty algorithms;
- **Rights and Liberties**: How can AI be used to support **authoritarianism** and either supporting or eroding citizens' rights and liberties in many domains;
- **Ethics and Governance**: How ethical codes could be developed in a time of political volatility

Similar questions apply in the military use of LAWS.
in "Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons" – a Briefing Paper for delegates at the Review Conference of the Convention on Certain Conventional Weapons (CCW) Geneva, 12-16 December 2016

Heather Roff argues that fully autonomous systems would violate current legal definitions of war that require human judgment in the **proportionate** use of force, and guard against targeting of civilians.

Furthermore, she argues that AI learning systems may make it difficult for **commanders to even know how their weapons will resp**ond in battle situations. Given these legal, ethical and design concerns, both researchers call for strict limitations on the use of AI in weapons systems.

Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control.
Autonomous Weapons: an Open Letter from AI & Robotics Researchers
http://tinyurl.com/awletter
announced at IJCAI Conference in 2015

Recommendations

ART in ARTificial Intelligence: Accountability, Responsibility, Transparency

A: To **whom** should we address if an autonomous vehicle runs over a pedestrian?

R: **System providers** are responsible for the **clarity** of the decision-making.

T: **System developers** must guarantee the right **specification**, development and deployment **good practices** of the AI systems.
To decide whether or not it is preferable to have the "Human in the Loop"

- **"The Human in the Loop"**
- **Privacy and data anonymization**.
- Appropriate Emotional states

Recommendations

- Precise **specifications** enforcing:

PRAGMATICS

  - "Intelligent" Systems **operation** to be clearly **understood**
  - including "**The Human in the Loop**" for critical decision-making, meaning **Accountability** in critical situations
  - Responsible **Data curators** guaranteeing DATA integrity
  - Systems (Robots/ Agents) architecture including "**Emotion-like**" states for human cooperation.

- **Legislation** for ethical principles in AI systems design

Actions are or have been done both at EU and USA

Say YES to some competences delegation!
Obsolete humans ? Never!

Societies are able to survive and rebuild after economic revolutions!
But many **concrete persons** may be crushed in the process
Our DUTY: do not allow that to happen

M. Delvaux proposed to the E. Union detailed legislation about civil rights and duties to be applied to AI entities, including intelligent robots limited **"e-personality".**
Law usually moves more slowly than Technology

How broader phenomena like **widening inequality** and intensification of concentrated geopolitical power and populist political movements will shape and be shaped by the development and application of AI technologies?
AI NOW report 2017

~We ALL should be civically interested in the future that is being shaped now

THANK YOU!

Eugenio Oliveira
(Email: eco@fe.up.pt)

Porto, December 6, 2017