Lesson 1_Castelfranchi

For a Science-oriented AI & not servant of the business

Introduction by professor Castelfranchi

Two main points

- **Why AI should not be reduced to a mere technology**
- **Why AI should care about economical, political and social problems underline the ethical ones**

Let's focus on the first point:

We (will) live in an AUGMENTED and MIXED world/"reality", not just "Onlife" on the WEB (Floridi's expression); living "connected", but in a **new material world/reality**.

We will act in the Virtual for changing the Real; and vice versa. We are "present" where we "are not"; we see and act where we "are not". And "somebody", which is not "here", will in fact act here and be "present" here.

We (will) live in a HYBRID Society, a mix of human intelligences and artificial ones, **not only Robots**, but **Intelligent software Agents** or Agents in our smart environments (house, office, cars,..) and **our cognitive prostheses**.

AI is not just building **a new technology** but **a new Socio-Cognitive-Technical System**, a new world and a new form of society, it is an **anthropological revolution**.

MY GOD, OH MY GOD, IT'S A PROBLEM! MY GOD... ("YOU HAD CONDIVIDED" Sartor)

You are social engineers, are you aware of that?

I will focus on

A) The importance of the SCIENCE side of AI;

B) some problems and dangers of the Digital Revolution and of the "mixed" (virtual and physical) reality and "hybrid" society (natural and artificial intelligences) we will live in.

A) For a Science-oriented AI

The pleasure of research (also in AI) should primarily be knowledge, discover, ideas, not just application and technology.

AI has a too strong “technological identity” more than a scientific identity.

AI provides conceptual and cognitive (formal) instruments for modeling and thus explaining minds, intelligences, action and interaction, emotions, organization, knowledge.

"AI should be proud of the crucial contribution it gave to the ***scientific revolution*** *in XX and XXI centuries due to* ***the impact of the Science of the Artificial on behavioral and social science"*** (Herbert Simon - one of the fathers of cognitive science and AI).

In science the economic/social/technical outcomes should mainly be “collateral/unintentional” effects, and not the main goal.

There must obviously be a research not generically K-oriented (knowledge oriented), but oriented to solve problems, but also in this “applied” research the priority is knowledge, understanding, explaining, modeling...

AI sometimes looks a bit perverted at the full service of business, for providing new market products: the new richness, the new industrial capital (Google, Amazon, etc.) are mainly based on AI products.

The scientific advantages of the artificial and synthetic approach that AI can provide to social sciences (to our minds and society) **is understanding more by building and simulating.**

(ISTC-CNR group exploited that in several doimains. On language, on autonomy, cooperation, sociality, trust, emotions, norms, power, etc.)

AI scientific models should be used for:

1) for **modeling/explaining human & natural Intelligences**; (Grosz on Conversation & shared plans, Ferrari’s cit. Winograd - ??? NON SONO RIUSCITA A TROVARE NULLA)

2) for emulating them;

3) for creating new intelligence and its theory (“General Intelligence” that is the concept of intelligence not only restricted to the human one, human intelligence is only a specific kind of the general intelligence.)

Philosophers frequently claim that what AI and cognitive scientists are doing is to **“anthropomorphize” machines** (which cannot in principle really have “mind”, “intelligence”, “intentions”, etc. but just “simulate” them). On this debate see for example Floridi and Sanders. BUT it is exactly the other way around: what we are doing is **to “de-anthropomorphize” such concepts**, making them **no longer “anthropocentric” but more general and abstract,** and more clear, formalized, and “operationalized”. No longer common-sense “words”.

AI mission is not just to take concepts and theories from human/social sciences and philosophy and brutally apply them. AI gives back a **crucial contribution to these sciences**, and not just a “technological” one, by changing concepts, models, and theories.

By focusing our efforts in this direction not only our environment and society will be hybrid and augmented, but also our brain and mind will be augmented, new cognitive power and new functions. **Our cognitive capabilities will not just be improved, but changed:**

> it is not only matter of “mnemonic functioning”, externalized memory, data access and processing, of “reading”, of “learning by doing”; it is also a real **evolution of our “social cognition”** in the Hybrid society.

In particular the WEB (the so called “Minds Online”) and also the virtual reality we are living in will empower:

> **“collective intelligence and problem-solving**”,
>
> “**collective sense-making**”,
>
> “**knowledge capital and sharing**”,
>
> “**creativity**”
>
> We will have a **new “embodiment of our cognitive representations**”, we will change our perception of **space**, **time**, **intelligence** and we will live in an extremely “**externalized/distributed cognition** and mind”

(P.Smart, R. Clowes, R. Heersmink: “Minds Online”, 2017)

B) The AI revolution: empowering whom?

We – AI & MAS community - are responsible for the introduction of “Agents” as which are **“autonomous” entities** (proactive, with initiative, with their own learning, reasoning, evolution, ..) that are also **“social”** (they are able to cooperate with human by following true **“norms”-** but also to violating them is some situations - and critically adopting our goals - **not just “executing” them-** with over-help (doing something more than what we ask), critical-help, etc.)

This concept of intelligent agents is a **correct and unavoidable mindset for a real “Intelligence” interacting with us and helpful.**

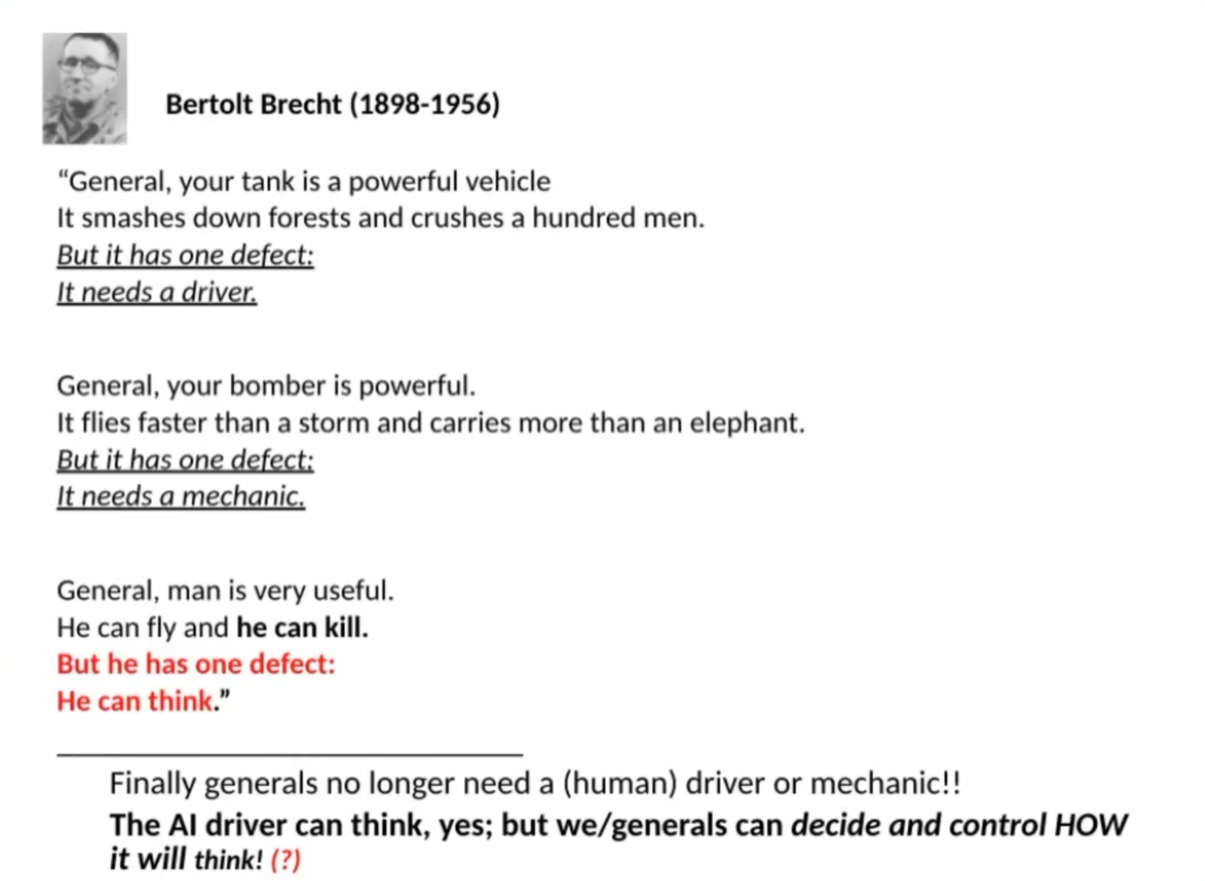
However, this model may have some problems, and scientists should **become aware** of **possible appropriation** of their creations, of possible **unacceptable uses** of their tools.

We have to ask ourselves: **are we missing the control?** Not only of our Autonomous Agents, Robots, etc., but of their **possible future (and present) uses.** Thus, the real question is:

Are we ready for the ANTHROPOLOGICAL REVOLUTION grounded on intelligence technologies and artificial mixed society? (Which is also an economic, social, and political revolution and not only an ethical one.)

Let's have a look at possible issues commonly identified by mass media:

- *Privacy*
- *Security (on WEB, … on access ..)*
- *Fake news, misinformation*
- *Hackers’ attacks*
- *Anthropomorphism*
- *War and Artificial soldiers/arms*
- *Ethics inside Artificial creatures and algorithms*



Bertolt Brecht (1898-1956)

“General, your tank is a powerful vehicle
It smashes down forests and crushes a hundred men.
But it has one defect:
It needs a driver.

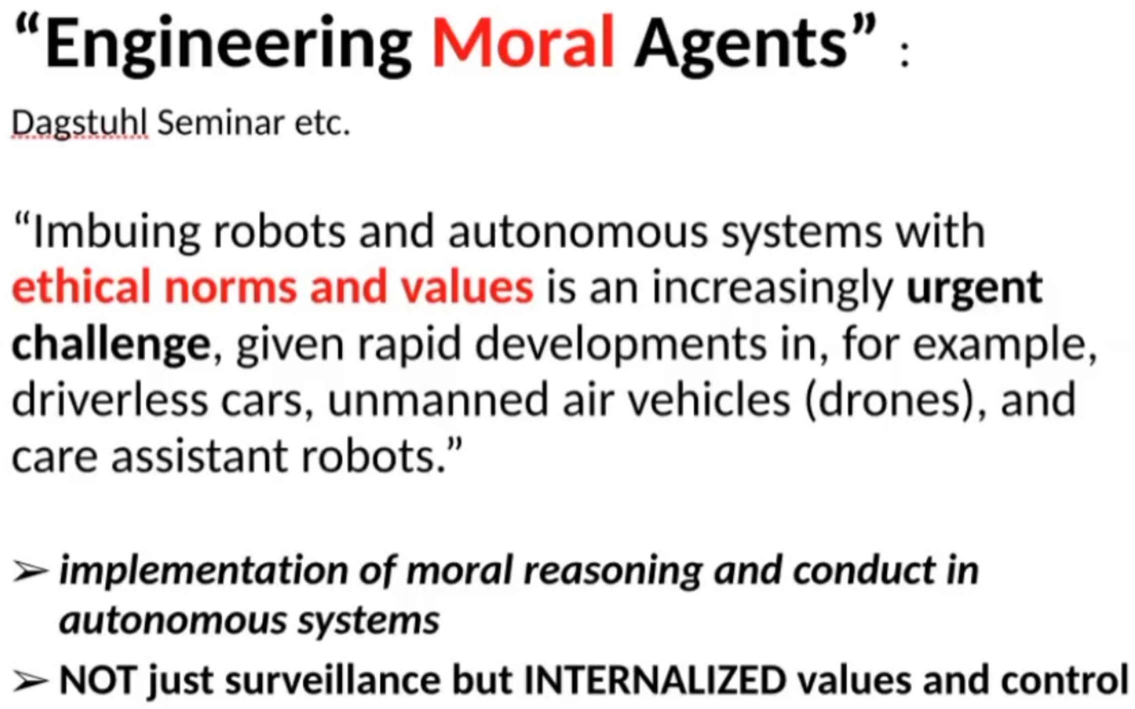
General, your bomber is powerful.
It flies faster than a storm and carries more than an elephant.
But it has one defect:
It needs a mechanic.

General, man is very useful.
He can fly and **he can kill.**
But he has one defect:
He can think.”

Finally generals no longer need a (human) driver or mechanic!!

The AI driver can think, yes; but we/generals can *decide and control HOW it will think!* *(?)*

“Engineering Moral Agents” - Dagstuhl Seminar 1622



“Engineering Moral Agents” :

Dagstuhl Seminar etc.

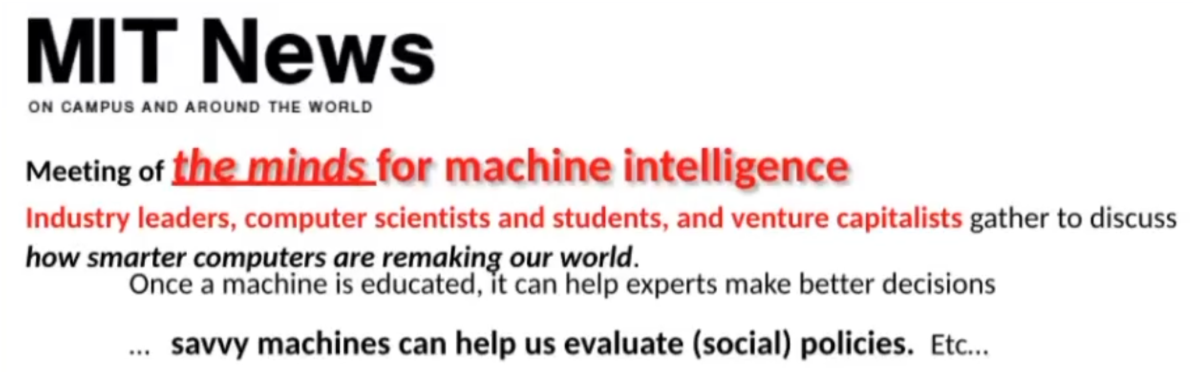
“Imbuing robots and autonomous systems with **ethical norms and values** is an increasingly **urgent challenge**, given rapid developments in, for example, driverless cars, unmanned air vehicles (drones), and care assistant robots.”

- ➢ ***implementation of moral reasoning and conduct in autonomous systems***
- ➢ **NOT just surveillance but INTERNALIZED values and control**

For Castelfranchi there are other not less serious problems related to the future of WORK in 4.0 economy!

B1) Is our Intelligent Technology research ONLY BUSINESS ORIENTED just because it needs money?

The questions in the boxes are Castelfranchi's questions.



MIT News

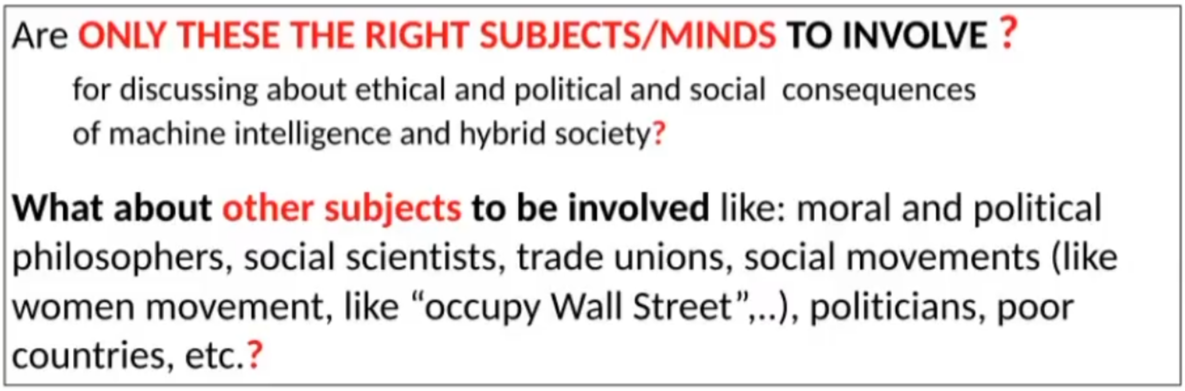
ON CAMPUS AND AROUND THE WORLD

Meeting of *the minds* for machine intelligence

Industry leaders, computer scientists and students, and venture capitalists gather to discuss *how smarter computers are remaking our world*.

Once a machine is educated, it can help experts make better decisions

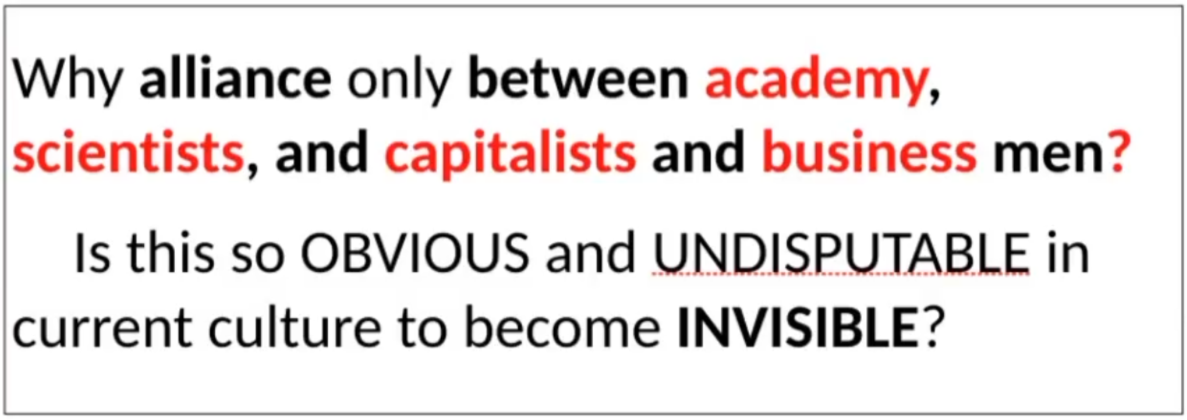
… **savvy machines can help us evaluate (social) policies.** Etc…



Are ONLY THESE THE RIGHT SUBJECTS/MINDS **TO INVOLVE** ?

for discussing about ethical and political and social consequences of machine intelligence and hybrid society?

What about other subjects **to be involved** like: moral and political philosophers, social scientists, trade unions, social movements (like women movement, like "occupy Wall Street",..), politicians, poor countries, etc.?

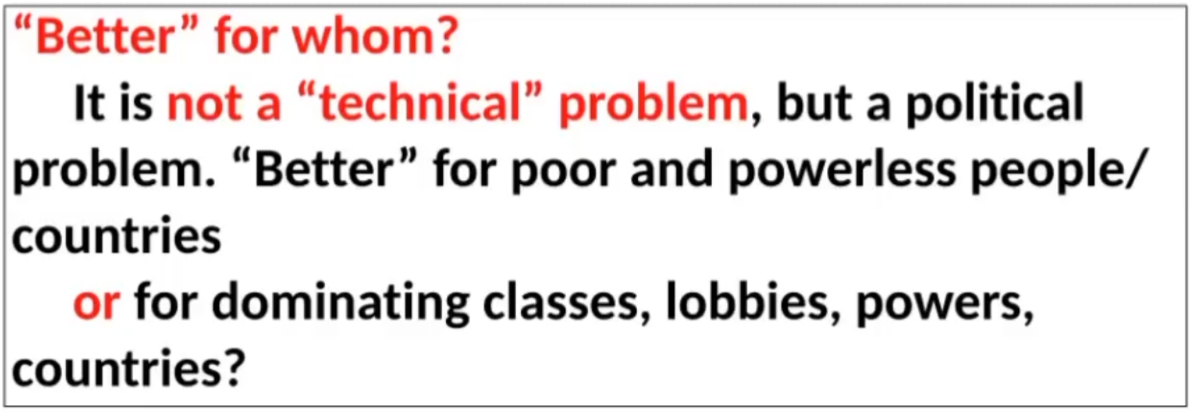


Why **alliance** only **between** academy, scientists, and capitalists and business men?

Is this so OBVIOUS and UNDISPUTABLE in current culture to become **INVISIBLE**?

In MIT conference they said that "*once a machine is educated it can help experts make better decisions, thus savvy machines can help us in evaluating social policies, etc.*"

Castelfranchi asked:



"Better" for whom?

It is not a "technical" problem, but a political problem. "Better" for poor and powerless people/ countries

or for dominating classes, lobbies, powers, countries?

Do not assume that *if something is beneficial it is beneficial for everybody*.

In society there are serious contrasts of interest and goals, thus if something is beneficial for X (that is favors his/her goals or interests) is noxious for Y.

For example

- If AI is subordinated to and beneficial for profit and business interests is NOT necessarily beneficial for workers.
- If is Beneficial for dominant countries not necessarily is beneficial for poor and colonialized countries.

For being **BENEFICIAL** AI should first choose on which side to be.

AI for sure can be very beneficial for

- democracy;
- good market, with reduced deception and manipulation;
- social planning and decision, and political imagination, projects;
- transparency and control, participation

But we need to carefully define it!

New Research Center to Explore **Ethics of Artificial Intelligence**

"AS A SOCIETY"?

"an array of academic, governmental and private efforts"

AGAIN:

- Why an alliance only between academy, scientists, and capitalists and business men, (and war powers)?
- Is this so OBVIOUS and UNDISPUTABLE in current culture?

- Is our Intelligent Technology research ONLY BUSINESS ORIENTED just because it needs money?

B2) Hidden Interests and Awareness Technology

By focusing our attention only on the ethical problems is a nice, but also very blind, approach. **We need to understand and investigate also the hidden interests behind some choices and decisions.**

Security, Privacy, War, Ethics are for sure very relevant issues that we have to reflect on, but they are not the only relevant ones from the *moral and political point of view*.

Hidden interests, manipulation of us (users and programmers), exploitation, emptying democracy, etc. are NOT less important.

Thus, scientists have to be conscious and not just manipulated or unaware although genial servants of those forces and interests.

In fact, **democracy is not a formal and misinformed voting ritual.**

We have to foster a **real "intelligence"** (understanding) and **EMPOWERMENT** of people in/on the hybrid societies evolution.

So we do not only need an **improved and collective INTELLIGENCE,** but also an **improved and collective AWARENESS,** which is a crucial form of "intelligence".

We need to understand what we are doing and why we are doing that to identify who is "nudging" us.

AI can help us in **RATIONAL DECISION MAKING** by revealing and correcting our irrational and affective **BIASES,** but the real goal should not to make "our" decision fully efficient and rational (not misinformed or biased), but to understand **in favor of whom we are taking such decisions.**

Intelligent Agents and algorithms have to help us to understand *not only our goals* and how to rationally decide (not misinformed or biased), but also to understand **in favor of whom. For example, which role are we playing in the society now?** We are consumers and we are not aware that we are constantly pushed by the system.

Thus, are the goals of our Agents and Robots they explicit, transparent at least to us?

(Ro)Bots & Agents should be **comprehensible** and trustworthy. They must be able to **EXPLAIN to us WHY they do/did what they do/did; The REASONS and MOTIVES of their actions, decisions, or suggestions.** NOT showing us their "algorithm"!

This requires a **COGNITIVE MODEL** of “reasons” and “motives” for believing, and for goal processing and decision (that is why **AI** should also be **for SCIENCE**).

What are the **INTERESTS**? **They are what is better for me and my goals, but I do not necessarily understand or intentionally pursue them.**

Tutelary Role (a crucial role in society):

> X takes care of my "interests”, of my good, even in conflict with me, with my current goals; X helps me or pushes me or obliges me! (E.g. mother/father/teacher ecc.)

In a lot of circumstances intelligent agents are deciding *for us* (delegated or not by us), or giving us recommendations or just a little push (the celebrated liberal “nudges”) *like in marketing.*

But **are they doing so in a TUTELARY ROLE?** (I.e. in my interests? Or in someone else interests?)

Brain storm di Castelfrancone (evidenziato in arancione):

- Who is judging what is better for me, or for us?
- Is this really “in *our* interest” or primarily in the INTEREST of financial and informational dominant powers? Or, as in many countries, of the political regime?

This holds also for more explicit influencing devices like ***RECCOMENDER SYSTEMS*** which should know us better then us.

- Will they give us recommendations and suggestions “in our INTEREST”, in a TUTELARY attitude, or will they follow market criteria just for a **more effective, personalized *advertising*?**
- Are they acting on the side of the “user”?!Or of the “seller” (of our *data* or of some *good*)?
- They will potentially **decide "for us**", but this expression is AMBIGUOUS: will they only deciding **“instead of”** us or also “**for our good”**?

Of course, **Social Robots** and **Intelligent Agents** will NOT govern in their own interest (science fiction!), but it is wise to ask in the interest of whom are they deciding?

- EMPOWERING whom?
- And will we be able to monitor and understand that?

- And to make that “transparent” to people?

Moreover, **“TUTELARY” doesn’t means only caring of our “individual” “personal” interests, but also helping us to understand and take care of:**

- **common interests and possible collective subjects and communities and pressures;**
- **hidden conflicts of interests;**
- **the “commons”, of public goods and their relevance and respect (environment, energy, water, public health, ...)**

“AUGMENTED INTELLIGENCE” also means AUGMENTED SOCIAL AWARENESS.

- How does the “INVISIBLE HAND” (the god of liberalism, which organizes the emergent and “**spontaneous social order**”) work?
- **Which political and moral values will the agents care of?**

B3) The "Mouth of Truth" Algorithm

Clearly, we are developing algorithms for ascertaining the “truth” in that mess of data, of assertions, hoaxes, and news, which are diffused and accessible through the WEB.

An Algorithm for deciding about *reliable sources, credible information*, what is “true” among so many different claims and data is necessary, there is no alternative on that.

However:

- On which base such algorithm will “ascertain what is true”?
- Only on the basis of reliable and convergent sources? Of their number and net? On direct or indirect access to the “original data”? Or also on the basis of the “values” and on the sharing and acceptability of the values of the source?
- These are crucial question also for the ‘official’ science: is it always capturing or saying the truth?
- And there will be dogmatic truths and undisputable authorities, like in any culture?

The people will believe what
the media tells them they believe.
- George Orwell

- And which culture and values will be assumed as the “right” ones?
- How will we allowed to distinguish between a *conflict of values or of interests* from a mere conflict between more or less credible data, more or less grounded, direct, controlled, reliable?

B4) 'Presences' in our Mixed Reality and Society

The **autonomous and proactive intelligent entities will become ‘presences’ and ‘roles’ in our *hybrid* society** (human and artificial agent) and mixed and augmented reality (combined *virtual and ‘real’*, *‘natural’* and *automatic/prosthesic* world).

Now the problem will be: **are we able to manage these *autonomous* and *too informed and intelligent* agents? And how will we relate to them?**

It is a matter of which roles those material or immaterial, visible and invisible “entities” will play in our life and environment.

- Will they be our Guardian angel with a ‘tutelary’ role? By helping, protecting and empowering us
- Or – less religiously – our Jiminy Cricket (The Talking Cricket) with its recommendations?
- Or our supervisor in the ICT-Panopticon we live in?
- Or our tempting Spirit
- Or our tempting Devil for the benefit of some marketing policy or monopoly, or the influencing and manipulating manager for hidden political or economic

powers?

Will a **MIXED REALITY include also a MIXED BODY & MIND??**

- Will we "incorporate", feel them as parts of "us", our "mental prosthesis"?
- Will we listen to that moral or rational "voice" as our own mental or consciousness voice our (expanded) *SuperEgo?*
- Or will our Super Ego be "externalized"? Not "me".
- Will we listen to "her" as to the voice of our mother, our teacher? Or will we become "voice hearers"??

!!!**NON HO CAPITO COSA INTENDE (lezione 1 - 1:11:00)!!!**

According to Castelfranchi, both the following solutions will be probably there:

The "social" one: externalized voices and Agents (Our best friend; our sexual partners)

The "reflexively social" one: an augmented internalized Self and Consciousness

It is a matter of which political and moral values they will care of.

B3) Disagreement Technologies

A)

Nowadays, **there is a too strong ideology and rhetoric about society as cooperation, collaboration, common intent, collective advantages,** how to reach convenient agreements and equilibrium, etc.

Moreover, the web is - non accidentally - favoring a deviating political feeling: *"we" against "them"* (governors, political caste, centralized powers). This perception of "we" is completely misleading: there is no a "we" with common values and goals and interests, which has to be unified against the political power as such (in case we should create a "we" against the real power - financial power - that has usurped the political power).

Population is composed of different classes, genders, generations, and cultures with very different and conflicting values and interests; this is the real conflict - not "we" and "them" - and political activity and forces *were* supposed to represent and protect those different social interests, and not just the "common" interest (in fact we have parties in politics).

Some conflicts of interest or of value can be solved and reconciled in a common interest, but a large part of political/government decision is not for a common advantage (except reducing civil war), but for a fair distribution; it is for the prevalence or advancement of the interests of a given group (class, lobby, gender, view, …) by reducing the powers of the others.

There is a need for conflicts since they are the presupposition of Democracy. There are not only conflicts of views or opinions, or due to different conceptions, information, reasoning, but also are conflicts of "objective interests" (of different groups or classes, private vs common interests.

All of these social conflicts **do not have a "verbal/cognitive" or a "technical" solution, merely based on data and technical principles, they have a "political" solution.**

Democracy is not only a "response" to conflicts or a way to moderate them; in fact, democracy actually **encourages the conflicts** and then tries to solve them, since **conflicts are the motor and principle of Democracy and of its possible effectiveness in changing society** in favor of the submitted subjects, disadvantaged classes and groups, etc.

The main problem of democracy is that we vote in a self-defeating way, and, in general, there is a collective stupidity. So, the question is:

- Might a (political) education to digital society and participatory democracy be enough to solve this "cognitive" and social problem?

According to Castelfranchi, they could help us, but he is a bit skeptical about the feasibility in a short period (maybe in a couple of centuries...).

However, **one of the main tasks of intelligent social technologies should be**

- to give voice to people that are not in the condition to protest, and to be listen to,
- to make conflicts *to* emerge and become aware of,
- *to* make express disagreement
- to making transparent which interests are hidden and prevailing
- ...

(B) "Critical Thinking"

Using WEB technologies for organizing "movements" it is fine, but it is not so good without promoting critical consciousness

We need environments and agents for **learning and developing a *"critical thinking"* attitude to manage our *cognitive and motivational biases.* They shouldn't be just used for selling and for dominating.**

Fo example, a problem is to **demystify the ideology of the NET:** *we perceive NET interaction as non hierarchical, without superstructure and mediation, individually managed, spontaneous, thus "free", really and directly "democratic" (one counts one).*

But this a wrong neoliberal view and perception: in fact, **there are new Powers beyond the WEB and its activity and information**; there are impressive oligopolistic economic interests, influence, manipulation and exploitation of data and work.

(C) Anti-manipulations

ICT and cognitive technologies are used for **recognize our profile and interests NOT for EMPOWERING US, but in order to propose/induce us to "buy" something** (goods, ideas, ..) They are monitoring and analyzing us in order to manipulate us and influence our choices.

Thus, we need *anti-manipulation* AI technologies:

> Castelfranchi would like to have a "l**ife navigator**" in my main "social role" (ex. consumer!), but not a navigator saying "turn right, turn left", "buy that; do not buy this"…but **a tutor, a trainer, inducing him to understand and to reflect about why he is oriented in that direction**, why he is choosing that product; worrying if he has the right information.
>
> Making him **conscious of who and how is persuading or just unconsciously manipulating him**; and so on.

Concluding Remarks: a glass of invisible

The great revolution of ICT, of digital *monitoring* and *predicting* (by simulation) and big data can give to society (and democracies) a glass were to observe themselves and follow what it is happening, see the hidden presences and observe the future (predictions and planning).

Goal-Oriented Agents Lab

The Goal-Oriented Agents Lab (GOAL) is an interdisciplinary group that carry out research on finalistic behavior in intelligent agents. Key areas of activity are Cognitive Systems, Social

http://www.istc.cnr.it/group/goal