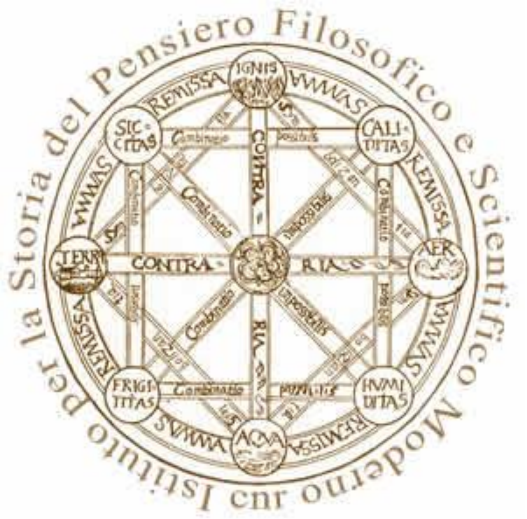
Cristiano Castelfranchi

For a Science-oriented, Socially Responsible, and Self-aware AI: beyond Ethical Issues



Laboratorio dell'ISPF, XIX, 2022

[4]
DOI: 10.12862/Lab22CSC

1. *Premise*

I will just to list and synthesize the main claims and arguments about what I mean by "responsible" and "self-aware" AI.

2. *Social Engineers*

We (will) live in an *augmented* and *mixed* world/"reality" (Ricci et al. 2015, 2017). 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.

Moreover, 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 mental prostheses*.

Not only our environment and society will by hybrid and augmented but our brain[1] 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, knowledge explosion; and "learning by (virtually) doing".

There will also be a serious evolution of our "*social* cognition" in the Hybrid society. In particular the WEB ("Minds on Line"; Smart et al. 2017) and Virtual reality will empower:

> "*collective* intelligence and problem-solving",
> "*collective* sense-making",
> "knowledge capital and sharing",
> "creativity",

and

> a *new "embodiment* of our cognitive representations": our perception of *space*, *time*, *intelligence*,… will be changed,
> an *extremely "externalized/distributed* cognition and mind".

In particular AI & MAS community are responsible for the introduction of "Agents" as "autonomous" (proactive, with initiative, with their own learning, reasoning, evolution, ..) and "social"; cooperating with human by following true "norms" (but also – in case – violating them), and critically adopting our goals (not just "executing"), with *over-help, critical-help*, … (Falcone e Castelfranchi, 1999).

[1] See for ex. Ienca 2019.

This was a correct and unavoidable solution, for a real "Intelligence" *socially* interacting with us and usable from humans.

However, this obliges scientists to become aware of possible appropriation of their creations, of possible unacceptable uses of these instruments:

Are we missing the control? Not of our Autonomous Agents, Robots, etc. but of their *possible uses*?

Are we ready for the *anthropological revolution* grounded on Intelligent Technologies and artificial mixed society? Which also is an *economic, social, and political revolution*.

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

Are "social engineers" aware this?

3. *For a Science-oriented AI*

AI has a too strong "technological identity" more than a science identity.

Actually AI provides *conceptual and cognitive (formal) instruments for modeling and thus understanding* 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 an social science (Herbert Simon)

There must obviously be research not generically K-oriented ("basic") 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. etc.)

The *scientific* advantages of the artificial, synthetic approach to mind and society is *understanding by building and simulating.*

AI scientific models:

1) for modeling/explaining human & natural Intelligences;
2) for emulating them;
3) for creating new intelligence and its theory ("General Intelligence").

What AI is doing and has to do is building an artificial and hybrid society, based on human and artificial "sociality", requiring "social" agents, that is "social minds" in relation with each other.

Philosophers frequently claim that what AI and cognitive scientists are doing is to "anthropomorphize" machines (that cannot in principle really have

"mind", "intelligence", "intentions", etc. but just "simulate" them).[2] 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 clearer, formalized, and "operationalized". No longer commonsense "words".

AI mission isn't just to acritically buy concepts and theories from human and social sciences and philosophy for "applying" them. It gives back a crucial contribution, not just "technological", by changing those concepts, models, and theories.

Moreover: how to build a trustworthy digital society with trustworthy artificial partners? Shouldn't "human centered" AI systems be conceived in this broader perspective, not only in terms of H-C-I, dependability, etc? And it is true that these artificial systems are "intelligent", "social", and can be normatively regulated; or they just simulate that?

4. *Alert for possible dangers (a limited view)*

Are there dangers in *living with* Artificially Intelligent Agents and Robots? Being replaced (practically or *cognitively*) or supported and guided by them?[3] Are there dangers in augmenting our intelligence and changing cognitive processing?

For the mass media (and also in our own debate), the main problems are: *Safety*, *Privacy*, *Security (on WEB, … on access ..)*, *Fake news, misinformation, Hackers' attacks*, *Anthropomorphism*, *War and Artificial soldiers/arms*, *Impacts on occupation/workers*, *Ethics inside Artificial creatures and algorithms,..*

And we have to work on

> Ethical issues, and
> for a Reliable and Transparent and Explainable AI.

This view is definitely important but for me limited and perhaps even hypocritical.

Aren't the exclusive *ethical focus useful blinders for covering the deepest problems*, like "Digital Capitalism" (for ex. Betancourt), dominant powers, etc.

5. *The AI Revolution: Empowering Whom?*

Ethics, security, privacy, war, technology transparency, trustworthiness, ... are for sure very relevant issues, we have to reflect on; however *not the most or the only* relevant ones from the *moral and political point of view*.

[2] On this debate see for example Floridi and Sanders.

[3] See for example the beautiful contents/topics of the Call for paper of "Robophilosophy Conference 2020: *Culturally Sustainable Social Robotics"*; with an important awareness of the ongoing "cultural" and social revolution.

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

For example the idea that: "The power is of the algorithms" (Stigler); "algorithms (will) decide for us" (see for example Beer, 2017). Is this fully true? Aren't there underlying interests, real "powers"? Don't the algorithms serve or not accidentally favor specific economic, social, political interests? Shouldn't this problem have the priority over the current debate just on the moral principles to be guaranteed in AI applications?

This is my question: political rather than ethical.

What is heard is wrong: that the future is in our hands and that the future of AI is in the hands of the creators of algorithms and robots, of the designers. It's in the hands of those who have power, command / decision, funding.

6. *Is AI Research only Business oriented?*

Is AI research aware of its use and orientation, or too servant of the business for its need for funding? This should be "transparent", not just the algorithm or the robot. Consider this prestigious meeting:

> **MIT 2018: "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 <u>better</u> decisions… savvy machines can help us evaluate (social) policies. Etc… (MIT News)

Two Questions:

> (A) 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.?
>
> (B) *"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?

As said, 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. We are "social engineers. Shouldn't *"human centered" AI systems* be conceived in this broader perspective, not only in terms of H-C-I, dependability, etc?

7. *AI for Freedom and Awareness Technologies*

AI can be very beneficial

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

"AI for FREEDOM" (JICAI-ECAI '18) is a great slogan! but *freedom of people*! not of dominant powers.

Just AI can provide *revolutionary instruments* for that: for making visible the "invisible hand" dominating market and society; for making transparent hidden alliances, interests; for making transparent the complex or hidden effects of collective behaviors; to make them more predictable by Dig Data and Simulation experiments.

AI can be a revolutionary "Awareness technology". It can not only improve personal and collective intelligence but collective *awareness*, which is a crucial form of "intelligence"; *understanding what we are doing and why we are doing that; who is "nudging" us.*

AI will help us in rational decision making, (by revealing and correcting our rational & affective biases); but… the real problem is not that "our" decision be fully efficient and rational (not misinformed or biased), but: in favor of whom? The *awareness* of "interests" we are serving.

"Augmented Intelligence" also means *augmented social awareness.*

How does it work the "invisible hand" (the god of liberalism) (Castelfranchi, 2014) which organizes the emergent and "spontaneous" social "order"? Can we show that, make it "transparent"? (§ VIII)

This holds also for more explicit *influencing devices* like *Recommender Systems, advisors,* which will know us better than us.

Will they give us recommendations and suggestions "in our interest", in a tutelary attitude, or will they follow market criteria with just a more effective, personalized *advertising*? On the side of the "user"?! Or of the "seller" (of our *data* or of some *good*)?

They will decide "for us", but this sentence is ambiguous: "instead of" us or also "for our good"?

Moreover we should/could Demystifying *the Ideology of the NET.*

NET interaction is perceived as *non hierarchical, without superstructure and mediation, individually managed, spontaneous, thus "free". Really and directly "democratic".* A neoliberal view and a wrong perception.

- There are new Powers beyond the WEB and its activity and information;
- Impressive oligopolistic economic interests;
- Influence, manipulation;
- Exploitation of MY data, Exploitation of MY work: can I "see" that?

We need *anti-manipulation* AI technologies.

As said, Social Robots and Intelligent Agents will not govern *in their own interest* (science fiction!) but… in the interest of whom? Empowering whom? And will we be able to monitor and understand that? And to *make that "transparent" to people?*

We need environments and Agents for learning and developing a *"critical thinking"* attitude. Not only to manage our *cognitive and motivational biases*; to support us in argumentation and discussion, and in understanding the tricky arguments of the others; or to resist to the prevalence of "audience" against "quality", of self-marketing and indexes against originality and quality; etc...

But also about propaganda, Academy, gender models, fanaticism, superstition, urban legends, …

We have impressive possibilities with new intelligent and interacting technology, big data, etc. They shouldn't be just used for selling and for dominating.

8. *The Mirror of the Invisible*

The great revolution of ICT, of digital *monitoring* and *predicting* (by simulation) and BIG DATA, can give to society (to demos) a glass were to observe themselves and follow what it is happening.

A mirror reflecting also what is invisible: presences and the future:

Not only hidden "presences"; what is "not present" here, but can be *virtually present for interaction*, and can act in this world and vice versa, etc.

But also a glass able to *show what cannot be seen/understood*: the future, predictions (for planning) the "emergent" order, and hidden phenomena and interests: for example, can I see who is now getting my personal data? And for what? For whom am I working for free?

Is sum: to see what Is (currently) invisible: Artificially Augmented Awareness is the real revolution of AI: Including itself! Its uses.

Shouldn't such awareness of real AI roles/uses also be an internal reflection and fight, not just philosophical and sociological monitoring and remarks from outside?

References

Beer, D., 2017. The Social Power of Algorithms, *Journal Information of Communication & Society*, Vol. 20, 1: "The Social Power of Algorithms".

Betancourt, M., 2016. The Critique of Digital Capitalism: An Analysis of the Political Economy of Digital Culture and Technology, punctum books. <https://doi.org/10.21983/P3.0125.1.00>.

Castelfranchi, C., 2014, Making Visible "The Invisible Hand". The Mission of Social Simulation. In D.F. Adamatti, G. Pereira Dimuro, H. Coelho (eds.) *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling*. IGI Global.

Falcone, R., and Castelfranchi, C., 1999. Levels of Delegation and Levels of Help for Agents with Adjustable Autonomy, *AAAI Spring Symposium on Agents with Adjustable Autonomy*, March 22-24, 1999, Stanford University, pp. 25-32.

Floridi, L., and Sanders, J.W., 2004. On the morality of artificial agents. *Minds and Machines* 14 (3), 349-379.

Ienca, M., 2019. *Intelligenza2. Per un'unione di intelligenza naturale e artificiale*. Rosenberg & Sellier.

Ricci, A., Tummolini, L. and Castelfranchi, C., 2015. The Mirror World: Preparing for Mixed-Reality Living. *Pervasive Computing*, 1536-1268/ IEEE.

Ricci, A., Tummolini, L., and Castelfranchi, C., 2017. Augmented societies with mirror worlds. *AI & Society*, 1-8.

Smart, P., Clowes, R., and Heersmink, R., 2017. Minds Online: The Interface between Web Science, Cognitive Science and the Philosophy of Mind. *Foundations and Trends in Web Science*, vol. 6, 1-2, 1-232.

Cristiano Castelfranchi
Istituto di Scienze e Tecnologie della Cognizione – CNR, Roma
cristiano.castelfranchi@istc.cnr.it

– For a Science-oriented, Socially Responsible, and Self-aware AI: beyond Ethical Issues

Citation standard:
CASTELFRANCHI, Cristiano. For a Science-oriented, Socially Responsible, and Self-aware AI: beyond Ethical Issues. Laboratorio dell'ISPF. 2022, vol. XIX [4]. DOI: 10.12862/Lab22CSC.

Online: 31.12.2022

ABSTRACT
Are we ready for the anthropological revolution grounded on Intelligent Technologies and artificial mixed society? Which also is an economic, social, and political revolution. AI is not just building a new technology but a new Socio-Cognitive-Technical System, a new world and a new form of society and culture. It is an anthropological revolution. Is our Intelligent Technology research only business oriented? AI should be more "science oriented". As for the possible dangers of AI impact, there is a dominant limited view, focused only on ethical issues, and on a reliable and transparent and explainable AI. My question is political not just ethical: AI revolution is empowering whom? AI can play a very important role "for freedom". It can also be a revolutionary "Awareness technology". It can improve not only personal and collective intelligence but collective awareness as well: understanding what we are doing and why we are doing it; who is "nudging" us.

KEYWORDS
AI impact; AI Ethical issues; Responsible AI

SOMMARIO
Per un'IA orientata alla scienza, socialmente responsabile e consapevole di sé: al di là delle questioni etiche. Siamo pronti per una rivoluzione antropologica fondata sulle tecnologie intelligenti e sulla società mista artificiale? Che è anche una rivoluzione economica, sociale e politica. L'IA sta costruendo non solo una nuova tecnologia, ma anche un nuovo sistema socio-cognitivo-tecnico, un nuovo mondo e una nuova forma di società e cultura. È una rivoluzione antropologica. La nostra ricerca sulle tecnologie intelligenti è orientata solo al business? L'IA dovrebbe essere più "orientata alla scienza". C'è una visione dominante limitata circa i possibili pericoli dell'impatto dell'IA, focalizzata soltanto sulle questioni etiche e sull'IA affidabile, trasparente e spiegabile. La mia domanda è politica, non solo etica: la rivoluzione dell'IA sta dando potere a chi? L'IA può svolgere un ruolo molto importante "per la libertà". Può anche essere una rivoluzionaria "tecnologia della consapevolezza". Può migliorare non solo l'intelligenza personale e collettiva, ma anche la consapevolezza collettiva: dobbiamo capire che cosa stiamo facendo e perché lo stiamo facendo; chi ci sta "spingendo".

PAROLE CHIAVE
Impatto dell'IA; Questioni etiche dell'IA; IA responsabile