

Review of EU Ethics & Trustworthy AI with Denning & Denning (2020)

A Review from the Point of View of the GCA Paradigm

*

Gerd Doeben-Henisch
doeben@fb2.fra-uas.de
Frankfurt University of Applied Sciences
Nibelungenplatz 1
D-60318 Frankfurt am Main

May 10, 2020

Abstract

The EU has published in 2019 a definition of AI¹ as well as a definition of Ethics for a trustworthy AI² in correspondence with this AI definition. In an acm article Denning & Denning (2020)[DD20] point to some dilemmas of AI. In this review it will be discussed whether and how one can deal with such questions within the GCA paradigm.

1 Trustworthy AI Ethics

This review starts with selected statements from the EU definition for an *ethic* dealing with *trustworthy AI* (cf. fig. 1).

Looking to all these statements *simultaneously* one can grasp some *general view* which sees *artificial intelligence (AI)* as a *process* which is *transformative* and *disruptive*. This implies that it is in general difficult to foreseen all possible outcomes of this process.

Because it is a clear position of the EU to *increase human flourishing*, thereby enhancing *individual and societal well-being* and the *common good*, as well as bringing *progress and innovation*, it is a strong intention of the EU to keep an

*Copyright 2019-2020 by eJournal uffmm.org, ISSN 2567-6458, Publication date: May-10, 2020

¹<https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>

²<https://ec.europa.eu/futurium/en/ai-alliance-consultation>

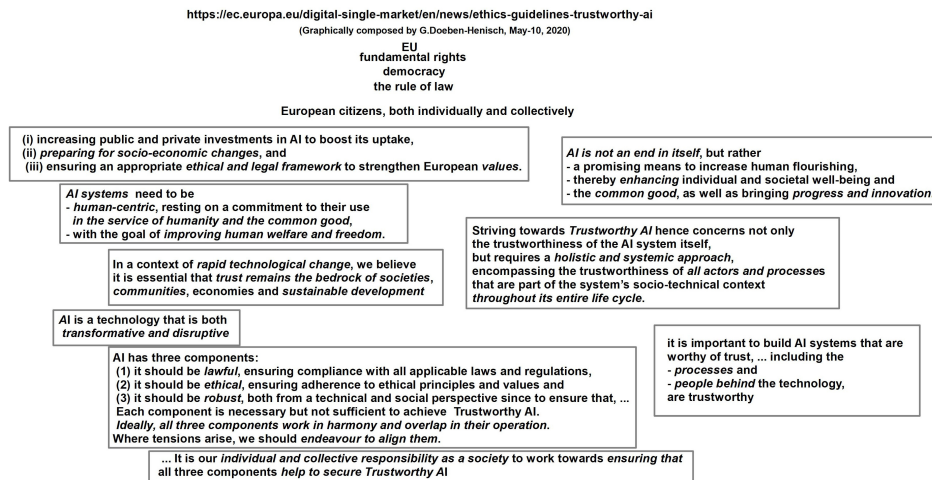


Figure 1: Some keywords from the EU definition for an ethics for trustworthy AI

eye on the development of AI to hinder negative consequences and to enhance the positive effects.

The only way to realize such an *ethical guarding role* is to give some *general criteria in advance* which should help individuals as well as the whole society to *watch* the developments of AI to detect negative effects as early as possible. Three main criteria are stated as follows:

1. AI should be *lawful*, ensuring compliance with all applicable laws and regulations,
2. it should be *ethical*, ensuring adherence to ethical principles and values and
3. it should be *robust*, both from a technical and social perspective ...

And it is commented further: "Each component is necessary but not sufficient to achieve Trustworthy AI. Ideally, all three components work in harmony and overlap in their operation. Where tensions arise, we should endeavor to align them."

It is assumed, that providing an AI fulfilling all these requirements that this can help to enable that trust which is the bedrock of societies, communities, economies and for a sustainable development.

2 Some Hidden Complexity

The clear and strong confession of the EU to enhance *individual and societal well-being*, the *common good*, as well as to bring *progress and innovation* points to some intended state in a yet *unknown future*. Even if we all would agree in these stated goals, the question(s), how to reach these goals in some given and *partially known present* are far from being clear.

From the point of a *Generative Cultural Anthropology [GCA]* the EU represents many Millions of actors (citizens), each with a very partial picture of the present world, and the past, and even more limited with regard to a possible future.

Taking this situation seriously we have to announce a list of several factors which contribute each to an *overall complexity* of this situation which makes the whole EU a challenge for everybody independent on which level of some hierarchies he/ she/ x is living and acting. Here are only a few of them.

2.1 So-Called AI

In the present there exists *real AI* not yet in the EU. The growing amount of algorithms and machines driven by such algorithms which can do some partial learning and organizing is far below those potential spaces of computation which are classified from many as *intelligent*.³

There exists some research to extend the potential of AI by enhancing the autonomy of the algorithms in the direction of a more *autonomous learning space* by extending the possible *values/ goals* which are guiding true *self-learning*. But as a recent review tell us⁴, the researchers cannot find any clue how such a *truly autonomous learning space* could look like. And looking to those actors which are known to be the *best actors actually known* to deal with values, ethic, and goals, the homo sapiens population, demonstrates since many thousand years that even homo sapiens, the humans, we, have not yet a clear concept for the generation of the needed goals for an unknown future.

In a recent paper Denning & Denning (2020)[DD20] have pointed out some problems which are given when dealing with AI, even with limited potentials of AI.⁵

³Despite the widely usage of the term *intelligence* this term is not yet sufficiently well defined to get acceptance in all related scientific and philosophical disciplines. The definition of the EU which is pointing back to the widely used text book from Russel & Norvig (2008)[RN10] represents only a very special view of AI!

⁴See Merrick (2017)[Mer17]

⁵And there are many similar papers around to discuss these topics.

1. **ANNs:** One kind of problems arises by the fact that a huge part of today so-called AI algorithms is realized by the usage of so-called *artificial neural networks [ANNs]*. The ANN-formalism has many strong points, but this strength is accompanied by critical points: (i) the formalism does not allow a transparent explanation *why* some results occur, if they occur. (ii) The final response of the system can be dependent from *small changes in the input*, which are for 'human experts' not really substantially. (iii) These AI-algorithms are in their final output highly dependable from the used *training sets*, which – somehow inevitably – are more or less *biased*. (iv) The preparation of *good training sets* requires time (which induces costs) and expertise which is beyond practical limits. All these factors are inherent in the ANN-technology.
2. **Fakes:** Digital technologies can meanwhile produce *artifacts* which cannot any more being distinguished in an everyday life situation from the *real facts*. In a world where we have to base our decisions on real facts this can cause failures and mistrust, the opposite of what the EU requires as ethical goal.
3. **Military:** The increasing usage of untrustworthy AI technologies within military applications without a public control increases the probability of harmful events in the future.
4. **Technological Singularity⁶:** Although the possibility of a technological singularity is not yet completely clear⁷, some authors classify the risk of a technological singularity realized by unlimited AI algorithms as high.
5. **Employment:** As already can be perceived today the potential of a change in jobs and business models, even in formats of whole societies, caused by still very simple AI-algorithms, is real and high. The used political and cultural techniques to react in face of these changes appear to be insufficient at a first glance.
6. **No more decisions?:** Delegating more and more decisions from humans to so-called AI-algorithms causes the final question about our role as humans on this planet: does there exist some important reason for us humans to be there, to be in command for the upcoming future *above the algorithms* or have we to state that there is only a *kind of nothing* driving us humans as part of the overall life on this planet and in this universe.

⁶For a first general overview see https://en.wikipedia.org/wiki/Technological_singularity

⁷We have already a real singularity given as the phenomenon of biological life on the planet earth

2.2 Safety Critical Systems

The discussions about trustworthy AI are usually centered around the perspective of software. But in the real world software never will occur as such; software is always embedded in some hardware thereby constituting a *technical system*.

There are many disciplines around dealing with technical systems due to different requirements which we put onto a technical system. In this context of interest are such disciplines which are looking to those technical systems which have a direct impact onto humans, for instance *real-time systems [RTS]* or – even more – *safety critical systems [SCS]*, systems whose behavior shall in no instance do some harm onto human persons.

It is Nancy Leveson, the 'master-mind' of SCS-theory for many decades, which has often very clearly pointed to those fundamental problems, which arise through the interaction of software and hardware in general. In a recent paper Leveson (2020)[Lev20] she summarizes these experiences in a condensed way.⁸

The central idea is the fundamental difference between the human thinking about the real world – a set of hypotheses, more or less partial, more or less vague and fuzzy – which encoded in symbolic structures as mathematical models and/ or algorithms (software) will be embedded in hardware which follows their own rules interacting with the physical world following their rules too. The *matching* between the *symbolically modeled* real world and the *real real* world can by principal reasons never be complete and therefore will ever include the potential for failures. The emergence of more algorithms cannot overcome this fundamental problem, but it can embed it in even more complex structures hiding errors in intricate ways.

3 The GCA Point of View

I will step back to the perspective mentioned above where the EU is described from the point of view of a GCA theory as representing many Millions of actors (citizens), each with a very partial picture of the present world, of the past, and even more limited with regard to a possible future.

As discussed in several papers of the *case-study section*⁹ of the uffmm.org site a *first fundamental pre-condition* for a common answer to given problems is the establishment of an *effective communication* enabling a *real cooperation*

⁸See my review of this paper here: <https://www.uffmm.org/wp-content/uploads/2019/06/review-leveson-2020-acm-yourSWillNotKill.pdf>

⁹<https://www.uffmm.org/2020/04/02/case-studies/>

between the different minds rooted in different brains. Such communication and cooperation requires besides several important *psychological factors* like e.g. 'trust' a sufficiently well elaborated *common knowledge*. Without such a commonly shared knowledge no real and good cooperation is possible. The *quality* of the resulting behavior and the accompanying *impact* of the society and the nature *directly depends* from the quality of this *available knowledge*.

Thus the main goal of a GCA theory is to clarify all conditions for a *good knowledge* as well as the conditions *how to reach* a good knowledge between as many as citizens possible.

The history so far shows that the big and successful cultures of the past are characterized by their more advanced structures of enabling *effective cooperation* as well as the enabling of a knowledge which is *better* than that of the competitors.

An important factor – but still difficult to understand and to handle – is *power*.

A *simple power model* is that which is based on concrete persons which are assumed to be the *representatives* of the power giving *directions* (*goals, values,...*) which way to follow in the future. This simple model has shown in the past, that it can work quite good compared to other systems if many supporting factors are in the right place; nevertheless these simple systems can also crash dramatically if the single person is crashing.

More advanced power models are relying on more complex systems of persons; one of the newest and rather modern systems are the *modern democracies* emerging in the 20th century.

But as we can observe today, these systems are not automatically better. The modern democratic systems presuppose not only a *sophisticated structure of power-management*, but even more a sophisticated system of *knowledge* presupposing an appropriate system of *education* and an *everyday value system* which is accepted by the *majority of the citizens*. This all demands further a very effective system of a *trustworthy public communication*. As soon as *only one* of these factors will not be good enough the system will encounter problems which can start a *negative feedback loop* to diminish increasingly the trust in the system.

References

- [DD20] Peter J. Denning and Dorothy E. Denning. Dilemmas of artificial intelligence. *Commun. ACM*, 63(3):22–24, February 2020.
- [Lev20] N.G. Leveson. Are you sure your software will not kill anyone? *Communications of the ACM*, 63(2):25 – 28, 2020. <https://doi.org/10.1145/3376127>.
- [Mer17] Kathryn Merrick. Value systems for developmental cognitive robotics: A survey. *Cognitive Systems Research*, 41:38 – 55, 2017.
- [RN10] Stuart Russel and Peter Norvig. *Artificial Intelligence. A Modern Approach*. Pearson Education, Inc. publishing as Prentice Hall, 3 edition, 2010.