Artificial Intelligence: ethical and social considerations

Francesco Corea*

ARTIFICIAL INTELLIGENCE: ETHICAL AND SOCIAL CONSIDERATIONS

ABSTRACT: Embedding ethical principles in the development of any technology is becoming more paramount as new questions arise on security, accountability, fairness and more. In this paper, we explained why the case for AI is different and call for better principles and thoughtful design. We then outline a set of recommendations that stem from a definition of rights resulting from principles and ethical values, and conclude with some brief discussion on biases and technical frameworks.

KEYWORDS: AI Knowledge Map; Bias; AI Principles; Data Ownership; Explainability

1. Introduction

There has been a lot of talk recently regarding the use (or misuse) of AI-systems in a spectrum of different scenarios (e.g., deep fakes, facial recognition, etc.), which are eventually bringing to light the importance for machines to be ethically designed. In fact, embedding ethical principles in our technology is an action we should not impose *a posteriori*, but rather in phase of design and development of the technology itself.

However, this is not the first time that academics, practitioners and policy-makers call to arms a specific industry asking for a better self-regulation. So why should it be different with AI and why it deserves so much attention?

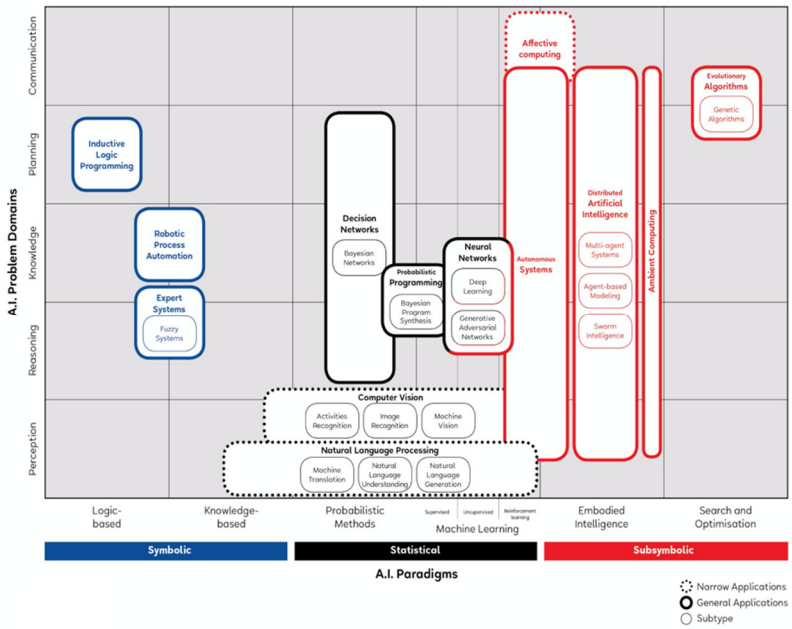
It seems that there are at least three reasons:

1) **AI is not only one thing**. Artificial intelligence is often identified with machine learning (and viceversa), as if there was a single unique component of this incredibly vast field of study. In reality, the space of AI is quite more cumbersome and complex, and consists of several tools and techniques. The AI Knowledge Map[1] is an attempt to map the diversity the field is made of, organize unstructured knowledge into a sort of ontology, and provide a gateway for researchers and specialists to tap into the disparate areas that compose the branch.

* *Independent Researcher, Email:* corea.fr@gmail.com .

[1] F. COREA, *An Introduction to Data*, Cham, 2019.

ISSN 2284-4503



On the axes, you will find two macro-groups, i.e., the AI Paradigms and the AI Problem Domains. The AI Paradigms (X-axis) are the approaches used by AI researchers to solve specific AI-related problems (it does include the approaches we are aware of up to date). On the other side, the AI Problem Domains (Y-axis) are historically the type of problems AI can solve. In some sense, it also indicates the potential capabilities of an AI technology.

In terms of AI paradigms, there are at least six there could have been identified[2] ;

1) **Logic-based tools**: tools that are used for knowledge representation and problem-solving;
2) **Knowledge-based tools**: tools based on ontologies and huge databases of notions, information, and rules;
3) **Probabilistic methods**: tools that allow agents to act in incomplete information scenarios;
4) **Machine learning**: tools that allow computers to learn from data;
5) **Embodied intelligence**: engineering toolbox, which assumes that a body (or at least a partial set of functions such as movement, perception, interaction, and visualization) is required for higher intelligence;
6) **Search and optimization**: tools that allow intelligently searching through many possible solutions.

In terms instead of problems AI has been used for, the classification used here is quite standard:

1) **Reasoning**: the capability to solve problems;
2) **Knowledge**: the ability to represent and understand the world;
3) **Planning**: the capability of setting and achieving goals;

[2] *Ibidem.*

ISSN 2284-4503

4) **Communication**: the ability to understand language and communicate;
5) **Perception**: the ability to transform raw sensorial inputs (e.g., images, sounds, etc.) into usable information.

The breadth and complexity of this map should send a clear signal on why regulating and infuse an ethical approach to the development of AI algorithms and applications is both so vital and hard. What it could work for neural networks (using machine learning to mainly solve reasoning/knowledge problems) could not apply for evolutionary algorithms (which use search and optimization method to mainly optimize for planning and communication).

2) AI affects our daily lives for real and intimately. From the algorithm that governs the recommendations for movie, songs or even products for e-commerce websites, to the one implemented to allow identity recognition and validation, to the one that controls your self-driving car or vacuum-cleaner robot, AI is hidden everywhere. But the key word here is "*hidden*". Often, in fact, consumers do not consciously make the choice of using AI (or nor they are aware the tool they are using is AI-driven), which makes the technology in question both easily-adoptable and dangerous at the same time. There are of course several fantastic applications that may help the world in ways we could not do before (e.g., AI for drug discovery, computer vision used to fight wild-animal poaching, machine learning used to detect and hinder cyber-attacks, AI used for scientific discovery or food molecular recomposition, etc.), but for each good application there is at least another bad we couldn't even anticipate (e.g., bots that go rogue, deep fakes and people impersonation, wrong matching and identification, etc.).

Hence, the potential of AI and the easiness with which it can be concealed make the ethical problem more relevant than ever.
But there is also another aspect, which is often not highlighted [3]. AI is so pervasive nowadays that interact with us at an intimate level, and slowly modify our behaviours and habits almost unnoticed. However, this diffusion-and-interaction aspect is actually two-fold: from one side, there is what is called "*paradigm 37-78*" [4]. We make machines better and they make us better off in turn. The paradigm is so-named after the famous Go challenge between Lee Sedol and AlphaGo. In the move 37, AlphaGo surprised Lee Sedol with a move that no human would have ever tried or seen coming, and thus it won the second game. Lee Sedol rethought about that game, getting used to that kind of move and building the habit of thinking with a new perspective. He started realizing (and trusting) that the move made by the machine was indeed superb, and in game four he surprised in turn AlphaGo at Move 78 with something that the machine would not expect any human to do.

[3] S. QUINTARELLI, F. COREA, C. G. FERRAUTO, F. FOSSA, A. LOREGGIA, S. SAPIENZA, *Intelligenza artificiale. Cos' è davvero, come funziona, che effetti avrà*, Torino, 2020, 153pp; L. FLORIDI, *What the Near Future of Artificial Intelligence Could Be*, in *Philosophy & Technology*, 32: 1-15, 2019.
[4] F. COREA, *An Introduction to Data*, cit.

From another hand, we tend to adapt more to the machines than they adapt to us (both at a personal level as much as the environment we live in). We change our habits and our environment to take advantage of the machines, which makes us prone to be manipulated by the creators of those technologies fairly easily and guided towards specific actions and outcomes.

3) **AI is a completely different technology wave.** If previous technology waves had at least one or two aspects in common, this is certainly not true for artificial intelligence. In fact, contrarily to the development of innovation such as the operating systems, mobile apps, cloud computing, web browsers, and much more closely to the breakthrough of the personal computers, AI is technology that has the power to be independently developed (many of the building blocks are open-source and easily accessible for people familiar with the industry tools) and modularly approachable (there is no need for developers and researchers to re-build the wheel every single time). This combination of independence and modularity fosters an almost unrestrained waive of innovation, because applications can be built at the edge in a fully decentralized fashion, and because the barriers to entry the field are not as high as the ones you would have, for example, in robotics.

Since the new advancements are not regulated (and nor are the applications spun out of those technical progress), the potential effects of the technology are gigantic (in both the directions) and therefore the call for ethical standards is incredibly urgent.

2. Principled Artificial Intelligence

The three reasons listed in the previous paragraph help making the point on why it is so important (but also complicated) to impose ethical design in the development of AI applications. This need has initiated in the last few years a stream of both literature in academia as much as industry frameworks that prevalently look at ethics and human rights as pillars of a fair ideation and usage of smart technologies [5] have tried to scholarly understand and summarize the main documents that have recently tried to address the issue, and identified a set of common observations. More in details, they uncovered a growing consensus around eight key thematic trends across forty-seven individual principles. Without going through all the principles, it could be useful anyway to have a look at the eight macro-areas:

1) **Privacy**: an AI system should respect individuals' privacy;
2) **Accountability**: the accountability for the impacts of AI systems should be appropriately distributed (and remedies provided);
3) **Safety and security**: AI systems should be safe, perform as intended, and secure from third-party attacks;
4) **Transparency and explainability**: it should be able to explain the decision of an AI system, as well as allow for oversight;

[5] J. FJELD, N. ACHTEN, H. HILLIGOSS, A. NAGY, M. SRIKUMAR, *Principled Artificial Intelligence: Mapping Consensus* in *Ethical and Rights-Based Approaches to Principles for AI*. Berkman Klein Center Research Publication No. 2020-1.

ISSN 2284-4503

5) **Fairness and non-discrimination**: AI systems should be designed and used to maximize fairness and inclusivity;
6) **Human control of technology**: important decisions should abide by human review;
7) **Professional responsibility**: professionalism and integrity should ensure appropriate stakeholders are consulted in the development of an AI-driven tool;
8) **Promotion of human values**: the end to which AI is built should be aligned with humanity's well-being.

Hence, starting from these major shared trends that seem to be commonly acknowledged and shared across all the different manifestos and frameworks, we have been able to form ourselves a more refined sets of Recommendations, derived from a list of Rights that result in turn from more general Principles and Ethical Values rooted in our social organization[6].

I) Principles and Ethical Values:
- *a. Human Dignity;*
- *b. Freedom and Civil Rights;*
- *c. Non-discrimination;*
- *d. Inclusiveness;*
- *e. Inequality Reduction;*
- *f. Social Cohesion;*
- *g. Damage Prevention;*
- *h. Peace and Justice;*
- *i. Sustainability.*

II) Rights:
- *a. Information;*
- *b. Education;*
- *c. Self-determination of Identity;*
- *d. Confidentiality;*
- *e. Protection of Rights;*
- *f. Rights of Weak Subjects.*

III) Recommendations:
- *a. Trust;*
- *b. Accessibility;*
- *c. Safety;*
- *d. Usability;*
- *e. Control;*
- *f. Responsibilities;*

[6] For full reference: S. QUINTARELLI, F. COREA, F. FOSSA, A. LOREGGIA, S. SAPIENZA, *Una prospettiva etica sull'Intelligenza Artificiale: princìpi, diritti e raccomandazioni*, in *BioLaw Journal*, 3: 183-204, 2019.

g. *Redress;*
h. *Data Ownership;*
i. *Governance;*
j. *Training.*

For the sake of brevity, we will not cover here the explanation of all Principles and Rights, but will jump straight to the Recommendations (which are very much in line with the majority of the documents Fjeld and colleagues assessed, but also propose some degree of novelty). So, first of all, it is paramount that technologies based on AI are reliable and trustworthy. Second, they should be transparent enough to be understood and explained, and so safe to guarantee both data privacy as much as personal safety (avoiding negative externalities and minimizing the incentives for bad actors to misuse the system). It should also be provided with an interface that facilitates the usability from human users, and supervised by human beings (to prevent the occurrence of unfair/unwanted decisions resulting from probabilistic computation). The responsibility dilemma still remains difficult to solve, but it is easy to imagine that for applications with significant societal effects an ex-ante responsibility (and accountability mechanism) should be set up. Moreover, in the same way we have the principle of "privacy by design" for personal data management systems, we make the case for a "redress by design" principle for AI systems (a principle that would allow for repair mechanisms in case of wrong outcomes generated by the machine itself). Also, the data pertains to the individual who generated them, while we should call as a society for a centralized authority that monitors, establishes a clear governance and regulates the dissemination of AI systems. Finally, we should design a training path that allows for a deeper understanding of the technology, but also would help requalification in case of job loss due to the introduction of automated systems.

3. Discussion and Conclusion

We have not consciously discussed in this paper hot topics such as autonomous weapons, human-in-the-loop (HITL) and human-on-the-loop (HOTL), but we want to cover a last very sensitive topic: data biases. AI systems inevitably inherit many of the biases from humans, and there are multiple ways they could be transmitted[7]:

1. **Data-driven bias**: the bias that depends on the input data used;
2. **Bias through interaction:** the bias that comes out from interactions with external parties that feed the system;
3. **Similarity bias**: it is simply the product of systems doing what they were designed to do (and that unintentionally restricts the possibilities of the system itself);

[7] K. HAMMOND, *5 unexpected sources of bias in artificial intelligence.* Retrieved from TechCrunch at https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/ on 5th Feb. 2021.

ISSN 2284-4503

4. **Conflicting goals bias**: the systems designed for very specific business purposes ends up having biases that are real but completely unforeseen (for example, see the paperclip-maximizer mental experiment in Bostrom, 2014)[8];
5. **Emergent bias**: the decisions made by systems aimed at personalization will end up creating bias "bubbles" around us.

Regardless of the source of the bias or the way it transmits to systems and decisions, it is fundamental for both developers and policy-makers to deeply familiarize with those and design mechanisms to reduce them. Even assuming a perfect data set, environment and learning process, we do not have any guarantee that at some point the AI system will not learn the same biases by itself, but this should not give us an excuse to not focus as much as we can on fighting biases that could exacerbates inequalities in phase of design and implementation of AI tools.

We want to also finish with a provocative thought: ethics is (partially) a technical problem. Some of the ethical concerns we have are essentially technical issues we are not able to optimize for, and in the same ways many of the potential solutions are merely technical (or could be technically explained). For absurd that it may seem, developing a technical framework to assess and correct AI systems may be an interesting first step toward more robust, safe, and trustworthy autonomous systems.

[8] N. BOSTROM, *Superintelligence: Paths, Dangers, Strategies*, Oxford, 2014.