Ethics and AI: the birth of an oxymoron?

By Fabrizio Cugia di Sant'Orsola

On Friday September 22nd 2023 the Future for Life Institute (FLI) released an *Open letter* calling for a six month pause on AI experiments, signed by more than 30,000 between world experts, researchers and top industry figures. The Open letter probably represents the most qualified request to shed a light on the principles of unregulated AI development, an attempt that follows the FLI prior publication in early 2023 on a long series of questions and ethical doubts still left unanswered on which ethical principles should direct AI service development. Naturally the Open letter again lists the series of possible threats on the protection of human rights which may stem from inadequate review of new AI services (particularly generative).

In 2019 the *Ethics Guidelines on AI* released by the High-Level Expert Group nominated by the EU Commission already stressed that "*AI requires not only compliance with the law: Laws are not always up to speed with technological developments, and can at times be out of step with ethical norms or may simply not be well suited to addressing certain issues. For AI systems to be trustworthy, they should hence also be ethical, ensuring alignment with ethical norms*." In view of such premise we now find a list of some preliminary "ethical-oriented" principles in the First draft EU Parliament AI Regulation (version released in 14.6.2023). For example the draft AI Regulation expressly bans AI social threatening systems based on cognitive behavioral manipulation, social scoring (a system presumably already adopted in China) and real time biometric recognition, this latter to be admitted only in case of specific judiciary guarantees granted to right-owners. Consistently the same draft Regulation establishes that generative AI (such as ChatGPT) should comply with transparency requirements, such as disclosing summaries of copyrighted data, content generated by AI and be designed to avoid illegal content origination.

With regards to threats on data treatment procedures, in March 2023 the Italian Privacy Authority imposed limitations on ChatGPT on the processing of Italian residents data in view of privacy concerns, addressing with immediate effect a provisional restriction on activities of data treatment on OpenAI, the US based company operating the platform.

What appears important to note is that the final draft of EU AI Act is expected to be adopted in 2024, and aside its transitional norms, all regulation should be deemed compulsory within the EU territory within the closing of 2026. Given the pace of technological developments, the term appears quite distant from meeting expectations and capable of responding to FLI enquiries. Development of AI services and related financing of projects is fostered at such a pace that services and solutions will be already at their second generation by 2026, such that imposing ex post ethical principles could simply appear unfeasible.

The conundrum brings to mind the philosopher Spinoza. Baruch Spinoza endorsed an expressivist conception of moral judgement: in his *Ethics*, stating stated that "*it is clear that we neither strive for, nor will, neither want, nor desire anything because we judge it to be good: on the contrary, we judge something to be good because we strive for it and desire it*". We would add that AI systems are well capable of generating their own market demand and easily overpower most ethical doubts on the basis of economic interests and market dynamics.

At its basics, the general ethical principles stated in the draft EU AI Act include: (i) Respect for human autonomy (ii) Prevention of harm (iii) Fairness (iv) Explicability. Many of such principles are to a large extent already reflected in existing legal requirements for which mandatory compliance is required and hence also fall within the scope of lawful AI, which is Trustworthy AI's first component. Yet while many legal obligations reflect ethical principles, adherence to ethical principles goes beyond formal compliance with existing laws.

For instance, with regards to the principle of respect for human autonomy the draft AI regulation suggests that AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans and instead should be designed to augment, complement and empower human cognitive, social and cultural skills.

The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight over work processes in AI systems.

With regards to protection of human dignity, AI systems and the environments in which they operate should be safe and secure and ensure that they are not open to malicious use. AI systems should also be designed as to even increase societal fairness, and provide equal opportunity in terms of access to education, goods, services and technology.

Yet our economy is profit-driven. Most ethical norms (such as fairness, non-discrimination, solidarity, justice mentioned in the draft Articles 21 and following), will be grounded on self-responsibility and accountability (explicability principle) which is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. Such cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights and ethical norms.

The EU draft AI Act sets on individual businesses the general "ethic by design" obligation in the development of AI systems, as businesses, not institutions, are interpreted to be the loci of moral decision making and moral responsibility in the EU economic system. Yet Ethics and AI could well be developing into a classic oxymoron: what appears clear is that AI systems stem from basic economic transaction-seeking and transaction-executing practices. AI development is leveraged on huge capital investments and follows business rules, and involves making representations to people (truth part) and relying on their representations (the trust part), consistently with the general business transaction definitions (cfr. Gini and Marcoux, "*The Ethics of business", 2012).* With the AI technological revolution at hand, entrepreneurs will not wait for customers to come in to them: they are already open to possibilities and imagine lifestyles that people may be attracted to, and if a design feature will be secured for the development of new AI services, this will most probably be market induction, demand self-generation.

In two words, there is the a need to set ethical rules on AI system development *prior* to the launch of services simply because AI represents a ground breaking technology capable of inverting the offer/demand principle.

Paraphrasing Wittggenstein ("*philosophy is not a theory, but an activity*") we may say that E*thics is not a theory, but an activity*. With regards to business ethics applied to AI we must first define and adopt on a world scale which ethics we intend to apply: a decision depending on the superiority of one ethical theory over the other is not one we can be confident of, and this controversy needs to be solved much before a choice needs to be done in practice, i.e. when a decision must be taken in business.

Context matters and we should not let the perfect be the enemy of the good. Joseph Schumpeter coined the phrase "*creative destruction*" to describe the dynamic effect of business competition. As in the case of genetic manipulation, sometimes technology develops to become antithetical to life and limitations must be set to a certain extent. In fact, a virtuous competitor harnesses the benefits of competition without allowing competition to become the master, and given its threats AI should certainly not represent the exception to the rule.