PONTIFICIA UNIVERSITAS LATERANENSIS

The Machine Ethics Challenge in AI

Autonomous AI System as Artificial Moral Agents

Gianfranco Basti

Faculty of Philosophy – Pontifical Lateran University – www.irafs.org

Summary

1. The ethical challenges of AI
2. AI systems as moral objects and subjects
3. Main ethical issues of AI systems as objects
4. Problems of AI systems as artificial moral subjects or agents (Machine Ethics)
5. The Machine Ethics (ME)
6. The distributed responsibility humans-machines: slow responsibility *versus* fast responsiveness
7. The opacity issue in AI systems and the AI «imitation game»
8. The Double Condition to Satisfy for a Consistent Attribution of an Accountable Moral Agency to AI Systems in ME

The ethical challenges of AI

- In the current **Age of Communication**, human beings and machines interact and depend on each other more and more strongly and inseparably, like as many ***conscious*** **and** ***unconscious*** **communication agents** (Basti, 2017).

- This paves the way for an increasingly broad and articulated discussion **on the ethical and legal implications** that the increasingly widespread use of AI systems in every field of everyday life of our social, economic and cultural life entails.

- As summarized very well in the essay on *Ethics of Artificial Intelligence and Robotics* in the *Stanford Encyclopedia of Philosophy* by Vincent C. Müller (Müller, 2021) – which I invite you to consult to have an in-depth, broad and updated picture to 2021 on the current debate – the discussion on ethics in AI can be divided **into two main strands**:

AI systems as moral objects and subjects

1. AI systems understood as **objects**, i.e. *tools* used by humans;
2. AI autonomous systems understood as **subjects**, that is, as *artificial moral agents*.

- Quoting directly from this article, the discussion on ethics in AI is currently concerning.:

- «Ethical issues that arise with AI systems as **objects**, that is, tools made and used by humans. This includes issues of privacy and manipulation, opacity and distortions, human-robot interaction, employment issues, and the effects of *autonomous* AI systems. Secondly, issues concerning AI systems as **subjects**, i.e. ethics for AI systems themselves considered as *artificial moral agents* in so-called **Machine Ethics**.».

Ethical issues of AI systems as objects

- First of all, let's see the ethical and legal problems concerning AI systems as **objects**, that is, as tools used by human beings (individuals, companies, public and private institutions, governments, ...).

- **Privacy and data manipulation issues.** They are the most obvious and easy to understand. AI systems are applied wherever there are large databases whose management is impossible for humans, and which now with the progressive digitalization of any aspect of personal, social and economic life, concern the sensitive data of all of us.
  - What perhaps escapes most and is paradoxical but true, is that these systems, profiling and crossing the data concerning us every time we use the internet or smartphone, make an online purchase, access a database, request an online document, or simply use a search engine on the internet, now know our habits, attitudes and preferences **much better than we know ourselves**.
  - And that these profiles are **accessible to others and not to ourselves** creates **a big ethical-legal problem** that we should sooner or later face as individuals and as governments.
  - Above all because **as of now** these profiles are **systematically used in the creation of fakes** to influence certain groups of people, with serious problems on the autonomy of choices not only in the economic-commercial field, but above all **in the political-social field**.
  - A **representative democracy like ours** no longer work if citizen choices are **systematically conditioned in a subtle but real way**. Hiding our heads in the sand as we are doing does not solve the problem but exacerbates it**.**

…Go on

- **Problems of opacity and distortion in data processing.** While the **AI-symbolic** (fully programmed) systems do not suffer from this kind of problems, the much more powerful **AI systems that include ML algorithms based on multilayer architectures of neural networks** (the so-called **deep learning**) systematically suffer from a problem of **opacity** to the same programmer in their decision process.
- **In expert systems of symbolic AI,** the inferential trees for data classification/manipulation are defined by the programmer and therefore the path followed by the system to reach the final decision is always controllable or **transparent**.
- **This is systematically impossible** in ML systems based on multilayer neural networks, which moreover necessarily **emphasize biases** or "negative propensities" towards certain groups – generally minorities – or types of individuals present in the statistical dataset on which **the training of the system is carried out** necessarily **biasing the decisions** of the system in a "non-transparent" or "opaque" way.
- Indeed, as we know, the updating of the statistical weights of the variables by back-propagation of the error in the supervised learning always takes place **blindly**, without taking into account the ethical relevance of the single variable involved, given that the system without the appropriate precautions, minimizes the overall error only with respect to the statistical distribution to be recognized / corresponded in input, also emphasizing the intrinsic distortions.

- All this with **serious consequences** when the data, predictions and decisions taken or suggested to the human operator by the AI system **concern the life, work, health, justice, or economic well-being** of people who instead **have the right to full "transparency" on the motivations** that determined the choices concerning them.

... Go on

3. **Human-robot interaction.** Although still not too evident to many compared to the previous problems, it is an emerging ethical-legal issue, which will become increasingly relevant, as robots and the AI systems that control them will be spread on a very large scale.

   - **Robots** – including autonomous aerial and ground vehicles – **are in fact destined to support or replace humans** in industry, communications (p.es., automatic call-centers), surgery (surgical robots), high-risk rescue operations and increasingly in military operations (armed drones, robot-soldiers, robotic artillery, etc.), all specific fields where they are already widespread. But also, in many other applications that affect the lives of all of us (think of **self-driving cars**), even the **most fragile**. Starting with **24/7 nursing**, **domestic care** and many other care relationships, even **educational** (distance teaching systems "intelligent" able to adapt to the individual student).
   - **Employment problems.** It is obvious that AI and robotics systems, to the extent that they replace humans in tasks related not only to manual and fatigue jobs as they were at the beginning but also related to services, **create employment and retraining problems of the workforce** on a large scale, including the need for compensation for workers who can no longer be retrained. And this has inevitable ethical-social-political implications on which governments must interven

“Algorithmic injustices” in ML systems

- «When machine learning systems that infer and predict individual behavior and action, based on **superficial statistical extrapolations**, are implemented in the social world, various unwanted but real problems arise. These systems emphasize and perpetuate social and historical stereotypes rather than profound causal explanations of these stereotypes. In the ML process, individuals and groups, often on the margins of society who fail to fit into “stereotypical boxes”, **suffer undesirable consequences**. Various results illustrate this: distortion in detecting skin color in pedestrians; bias in predictive crime policing systems; gender bias and discrimination in economic advertisements for STEM careers; racial bias in criminal recidivism algorithms; biases in search engines; prejudice and discrimination in medicine; and prejudices in hiring, to name just a few» (Birhane & Cummins, 2019, p. 1).

- This means that AI machine learning algorithms, when applied to automated support for decision-making in different social, political and economic spheres, **are not value-free or a-moral at all**.

Problems of AI systems as artificial moral subjects or agents (Machine Ethics)

1. **Decision-making autonomy of AI systems.** It is certainly the most ethically delicate problem of AI systems and of robotics, especially in those systems such as **self-driving vehicles** both on land and in air, particularly those used as weapons, or as robots with military applications (**autonomous weapon systems AWS**), medical and personal care, or finally – on an even wider application scale – such as home automation systems (**domotics**) .
   - Finally, what escapes public opinion is that there is a field in which the autonomy of AI systems is **applied on a large scale, at least from 2008 onwards**, in the aftermath of the great crisis of the financial markets on a global scale.
   - That of **automated fast transactions on the financial worldwide markets** that now cover more than 50% of the transactions themselves on the world equity, financial and commodity markets. The ML algorithms that are used in these AI applications **follow a logic of maximizing profits** without almost ever using any ethical constraint.

The Machine Ethics

2. **Machine Ethics.** It is clear, says Müller, that with this class of issues we are faced with the consideration of **AI systems as subjects** based on the simple syllogism that "if machines act in ethically relevant ways, then we need an ethics for machines or **Machine Ethics** (ME)".

♦ Generally,

- "Machine ethics is concerned with ensuring that the behavior of machines towards human users, and perhaps other machines as well, is **ethically acceptable**." (Anderson & Leigh Anderson, 2007, p. 15)

- "The 'reasoning' of autonomous AI systems should be able to take into account **social values and moral and ethical considerations**; weigh the respective priorities of the values held by the different stakeholders in various multicultural contexts; **explain the decision-making process; and ensure transparency** (Dignum, 2018, pp. 1-2)».

- "There is a broad consensus that **accountability, and liability** to moral and legal rules are fundamental requirements that must be respected by new technologies. (*European Commission, Directorate-General for Research and Innovation, Unit RTD.01*, 2018, p. 18)) but the question in the case of robots and AI systems in particular the autonomous ones is **how this can be done and how moral and legal responsibility can be assigned to machines"**

The distributed responsibility humans-machines: slow responsibility *versus* fast responsiveness

- "The effects of decisions or actions based on AI are often the result of countless interactions among many actors, including designers, developers, users, software, and hardware. This is known as distributed agency. With distributed agency comes *distributed responsibility"* (Floridi & Taddeo 2016).

- As suggested elsewhere (Basti 2017; Basti & Vitiello 2013; 2023), to solve this typical conundrum of the "shared responsibility" humans-machines, the extension to AI systems of the **neuroethical distinction** between the **fast unconscious responsiveness** of our brains to environmental constraints and the **slow conscious responsibility** of our minds becomes essential, given that AI systems are very much faster than our brains in taking decisions.

- "It is characteristic of conscious processes **that they are much slower than nonconscious***;* **the rapid responsiveness of highly skilled agents** like (…) Senna must certainly be driven by the latter and not the former. ***It therefore seems false that agents must be conscious of the information they respond to in order to be responsible for how they respond to it****.* (…) Direct moral responsibility requires that a creature conscious agent *be conscious of the moral significance of their actions"* (Levy

The opacity issue in AI systems and the AI «imitation game»

- What in his book Levy emphasizes is that in the human production both of **cognitive and of moral judgements** – formally, logical valuations/decisions “true/false” (1/0), in **alethic and deontic modal logics**, respectively – what is “transparent”, i.e., **conscious** to us and eventually **transparent** and then **accountable** to others, is what is **before and after** the production of the judgement (decision) itself that as such **is absolutely *unconscious* and then “opaque”** to everybody, just as it happens **to autonomous AI systems endowed with deep ML.** Indeed, in producing a moral judgement/decision:

1. At first, we ***consciously* examine** the different components of the action/choice that we are going to evaluate by a moral judgement over it, i.e., the past similar situations, the actual concrete situation, the future practical consequences of our action/choice, and of course also **the abstract moral norms** that should rule our action.

2. Afterward, by combining **through an *unconscious process*** these and other components not considered at the first step (mainly our **emotions**) we produce our moral judgement/decision. However, **being truly responsible of the moral significance** of our actions requires that, before executing our action/choice,

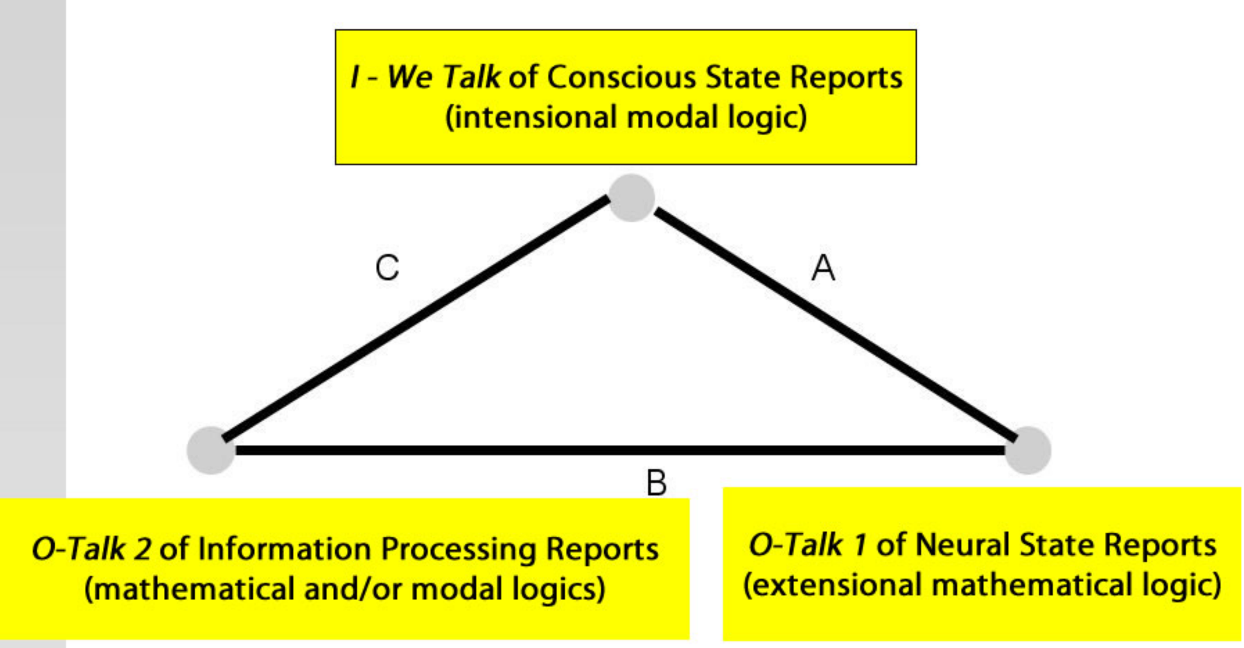
3. As a third step, **we make *consciously* a sort of “moral auditing to ourselves”** about our moral judgement (i.e., we perform **a moral higher order reasoning**) for evaluating whether effectively this judgement/decision we produced **satisfies all the moral constraints** we imposed before to produce our judgement/decision – and eventually other moral constraints we did not consider.

The imitation game and the cognitive triangulation in AI systems

- 1. **The *I/We-talk*** of the subjective intentional state reports in “singular/plural first person”. They are formalized in the “intensional logics” like as many (“ontic”, “epis-temic”, “deontic”) interpretations of the *modal calculus.*

- 2. **The *O-talk*₁** observational language of *neuroscience*, formalized in the *extensional logic* of the neuroscience mathematical models.

- 3. **The *O-talk*₂** of the observational language of the *information processing* in the brain. They can be developed, either in terms of the mathematical calculus of the *extensional logic*, or in terms of the modal calculus of the *intensional logics.*


I - We Talk of Conscious State Reports (intensional modal logic)

O-Talk 2 of Information Processing Reports (mathematical and/or modal logics)

O-Talk 1 of Neural State Reports (extensional mathematical logic)

The Double Condition to Satisfy for a Consistent Attribution of an Accountable Moral Agency to AI Systems in Machine Ethics: I

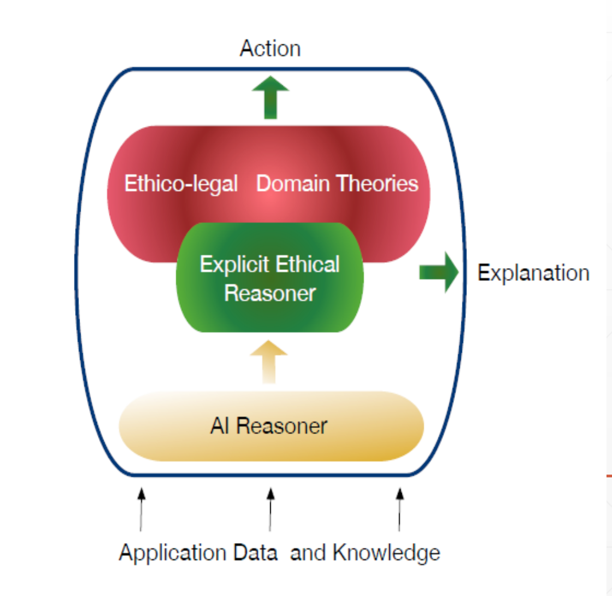
1. **The "transparent" implementation in the supervised ML algorithms of *ethical/legal constraints***, that is, error minimization algorithms satisfying ethical conditions (Lo Piano, 2020). In this sense, the so-called **"consequentialist" or "value based" approach to deontic logic** – i.e., formally satisfying the following modal logic scheme: "if you *want* to pursue this goal (value), you *must* do this" – seems to be more suitable for being directly implemented in ML algorithms since in both cases a cost function is to be minimized (Floridi et al., 2018) than the so called "virtue ethics" (Vallor 2017).

   - For instance, in the case of AI autonomous systems for automatic trading in the financial markets, an "ethically good" ML algorithm for trading means that it is not based only **on the maximization of profit**, but also **on the satisfaction of given ethical clauses** (e.g., investments not deriving from illegal origins, not based on the exploitation of the workers, etc.).
   - Finally, the value-based deontic logic is compliant also with **the implementation of "fairness conditions" in the data pre-processing by an *unsupervised* ML**, for avoiding the unwanted "bias" in the training data set of supervised ML (Gajane & Pechenikiy, 2018; Card & Smith 2020; Basti & Vitiello, 2023).

The Double Condition to Satisfy for a Consistent Attribution of an Accountable Moral Agency to AI Systems in Machine Ethics: II

2. The implementation in autonomous AI systems **of *an automatic ethical/legal auditing* to check in a transparent way** whether the decisions taken by **the system effectively meet the ethical criteria** transparently set in the ML algorithm and/or, in the case of symbolic AI systems, the ethical criteria implemented in the decision tree of the program.

   - Only recently the researchers in AI started to study this fundamental component of ME, **requiring a deontic HOL for a metalogical valuation** of the effectiveness of the deontic logic algorithms implemented in the ML program and/or in the inferential tree of symbolic AI systems (Benzmüller, Parenta, van der Torre, 2020).


Action
Ethico-legal Domain Theories
Explicit Ethical Reasoner
Explanation
AI Reasoner
Application Data and Knowledge