Ethics: names and what they affirm

ethics of AI is the branch of the ethics of technology specific to AI systems. It also includes the issue of a possible singularity due to superintelligent AI.

It studies moral behavior of humans and behaviour of machines

Rationality→ Ai achieves it by perceiving the environment reasoning on what is perceived deciding what’s the best action

Weak AI→ program that doesn’t have consciousness it is programmed to do specific tasks Narrow AI → ability of an intelligent agent to LEARN and perform a specific task with human proficiency

Agency: capacity of ai to act in a moral sense. Moral Agent: a being that can reason about morality.

Patient: AI systems are patients in ethical terms. Moral patients are those individuals who lack, either by age, physical or mental condition, etc, the ability to self-legislate or rational autonomy are moral patients. For example, a year-old child lacks the ability to engage in rational, self-legislating behaviour. The child is a moral patient.

To sum up: ethics of AI is not just about AI but it is also about ourselves too, AI functions as a mirror for reflecting on us

AI definition according to the European Commission High-Level Expert Group on AI Software systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected data, reasoning on the knowledge or processing the information derived from this data and decide the best actions to take. The issue is: what do we mean by the best actions, how do we decide?

Other definitions given by the commision

Rationality→AI achieves it by perceiving the environment in which the system is immersed through some sensors, reasoning on what is perceived, deciding what the best action is and then acting accordingly through some actuators thus possibly modifying the environment.

**Bartneck:**

Intelligent agent, what does it mean? First of all an agent is something/someone that acts, intelligent means that:

- Its actions are appropriate to its goals

- It is flexible to changing environments and changing goals

- It learn from experience

- It makes appropriate choices given its perceptual and computational limitations

AI can be divided into:

- narrow/ weak AI→ it performs only in specific domains

- strong/general AI→ capable of carrying out any cognitive task that humans can do

**Van De Poel:**

Morality→ totality of opinions, decisions and actions with which people express what they think is good or right. It is the totality of norms and values that actually exist in society Ethics→ is a systematic reflection on what is moral, it is not a manual with answers, it reflects on questions and arguments concerning the moral choices people can make.

Argument→ its aim is to justify or refute a statement, in general it can be expressed as: A1, A2, A3 so B (where An are premises and B is the conclusion). Argumentation is an activity that defines/ defends or attacks an opinion.

Descriptive ethics→ it describes existing moralities such as customs, habits, opinions about good and evil, acceptable and unacceptable actions… Description of existing morality Descriptive judgement→ it describes norms, it is either true or false

Normative ethics→ branch of ethics that judges morality and tries to formulate normative recommendations about how to act or live. Judgement on morality

Normative judgement→ it tells us how we should behave, it is good or bad

Values→ convictions that people have and think should be followed by everyone in order to be able to lead a good life.

Distinction between *intrinsic* and *instrumental* values:

Ethics deals with objective values, they’re also called intrinsic values, we can define them as aims we have to reach, they’re objective values in and of themselves. Instrumental values instead are means to realising an intrinsic value.

Exa. a person can think about its work as an intrinsic value

3 kinds:

Intended: designed by designers

Realized: technical artefact decides them by itself

Embodied : potential to realize a value in an appropriate context

Norms→ rules that prescribe what concrete actions are required, permitted or forbidden. These are rules and agreements on how people should behave and how they should treat each other. Values are translated into rules in order to make it clear ho we should realise values

Technical artefact→ it has a dual nature, it is an hybrid object with physical and functional features. The first ones are intrinsic features while functional features are related to its physical features on the one side, and they're also related to human intentions/practices of intentional human actions on the other side.

Nb it is just an abstraction, we don’t have it in real life

Technical artefacts vs mere physical objects: they differentiate in terms of their relational/ extrinsic properties.

*AI sistems as non traditional sociotechnical systems:*

AI systems are not just technical artefacts, they’re sociotechnical systems, this means that they don’t depend only on a technical hardware but also human behaviour and social institutions in order to function in a proper way.

Traditional sociotechnical systems consist of three basic building blocks: - Technological artefacts

- Human agents

- Institutional rules

AI system building blocks:

- Technical artefacts

- Artificial agents

- Technical norms (We create technical norms in 2 different ways:

1. Offline way→ norms are encoded by system designers

2. Autonomously discover→ specified and encoded in the agents by the (human) system designers. In the second case, agents may pick up norms from their environment in various ways, or norms may emerge from agents’ mutual interactions or interactions with human agents.)

Example: traffic lights

*Embodiment of values:*

We can design moral values into technical artefacts and so make them morally value-laden. Now the question is: which values?

For an entity to embody a value, there must be reasons for a pro- attitude or pro-behavior toward that entity.

Van de Poel defines a value in these terms: if x is valuable then one has reasons for a positive response (a pro-attitude or pro- behaviour) towards x, whereas the reasons originated from the object in itself.

In AI we define 3 different types of values:

1. Intended→ values the designer put inside technical artefacts

2. Realised→ values a technical artefact makes by itself depending on the context in which it operates

3. Embodied→ potential to realise a value in an appropriate context. An embodied value results from the properties that are designed into an object.

Two conditions must be met for technological artefact x to embody

value V:

1. x is designed for V;

2. (The use of) x is conducive to V.

Interactions between human agents are regulated via social institutions, we need to do the same thing for artificial agents so we translate such social institutions into computer code. Such computer code is made to give artificial agents technical norms.

According to Van de Poel, technical norm N embodies value V if (1) N has been designed (by the human system designers) for V and (2) the execution of N within the system is conducive to V”

What distinguishes us (human agents! Not only human because children are patients) from artificial agents(different from technical artefacts since they have autonomy, adaptivity and interactivity)?

1. AAs can embody values while humans can’t

2. AAs can’t embed values, they are the objects in which humans embed values 3. AAs and human agents can both adapt behaviours from external circumstances

**Floridi:**

*AI properties as sociotechnical systems:*

- Interactivity→ an agent and the environment can act upon each other they influence each other

- Autonomy→ the agent is able to change its state without direct response interactions, it can perform internal transitions to change its state. We can say that the system is autonomous since its behaviour is non-deterministic.

1. Personal autonomy refers to the capacity to form personal values, goals… 2. Moral autonomy means the possibility of reflecting on one’s own moral principles or ethical convictions

3. Rational autonomy is based on the weightiest reasons

4. Agential autonomy is the artificial systems’ ability to change internal states without external stimuli. It consists of identifying a certain behaviour as a ‘genuine action’, that is an action not entirely determined by external factors

- Adaptability→ the agent’s interactions can change the transition rules by which it changes state, it ensures that an agent might be viewed as learning its own mode of operation in a way which depends on its experience

To sum up:

Technical artefact→ physical and functional features. They’re physical objects designed by humans that have both a function and a use plane

1. Physical f are intrinsic f

2. Functional f are related to its intrinsic physical features and they’re related to human intentions or practices of intentional human actions

In AI physical features are translated into physical and causal terms

**Coecklebergh:**

*What’s ethics about?*

AI Ethics→ ethical challenges posed by current and near-future AI and its impact on our society. It is about human interactions, it is not only about technology but also about what humans do with it, how they use, perceive and experience it.

*Biases:*

Example: An AI is trained for job recruitment

They can arise from different situations:

1. All stages of design, testing and application

2. In the training set

3. In the algorithm

4. In the dataset

5. In decisions made by fake correlations exa. Nicolas cages’ films and drawings in swimming pools

How to solve the problem? Hypothesis: *The mirror view*

It means we use a data set that mirrors the real world, one could argue that this data set will represent prejudices inside and so the algo may still have biases hence we should not use such a dataset.

(Brief parenthesis

*Human-technologies interactions:*

There are 2 theories:

- Instrumentalist theory (**Gunkel**)→ technology is just a tool. It is not bad or good in itself but it depends on our use it’s neutral. Observation: it’s the use of technology that tells us whether an action we perform with it is good or bad. Criticism: take as example guns they’re designed only to kill and we can say that the world changes its

shape when we have one of them in hands.

- Mediation theory (**Ihde**) → technologies aren’t only an instrument but they also have a mediatic role. We shape technologies and technologies shape us. Technology isn’t neutral.

We have 4 types of human-technology interactions:

1. Embodiment relations → we use technology but we aren’t aware of using it. Exa when driving a car or using glasses *(I-tech)-world.* (**Verbeek** *Philosophy of Mediation*: obstetric ultrasound. Humans’ intentionality is mediated by technical artefacts)

2. Hermeneutics→ we interpret the world using technology. Exa the clock *I-(tech-world) (Coeckleberg)*

3. Alterity relations→ we are interacting with something other that exists but doesn’t influence the way we see or interact with the world. Exa ps4 or robots *I-tech-(world) (Coeckleberg)*

4. Background→ technology structures and mediates our experience but it remains in the background. Exa thermostat. )

*Problem with morality:*

While talking about human technologies relations, we defined four of them (see Ihde), the third, that is the alterity relation is given by Coeckelbergh. He proposes to build quasi moral

robots; we see them in a similar way as we do when another human being or a pet moves about in our environment and interacts with us.

Going back to the problem of morality, such robots produce the appearance of being moral, they should imitate subjectivity and consciousness in a convincing way, that is what humans do. Humans, in fact, don’t have the proof that another person has mental states and consciousness, they just interpret the others’ behaviours as emotions.

*Relational approach:* moral standing isn’t an objective property that we define just by judging the behaviour of others, we also need discussion - in a human language thinking way- to interpret in a more complete way moral states.

This means that robots might appear in different ways to different people in different situations, hence ethical attention passes *from ontology*, that deals with properties of an entity, to epistemology, that deals with our knowledge about the entity. It also passes from object to subject and from what things really are to how we look at them.

**Wallach and Allen:**

*Moral status of AI:*

While talking about moral status we have to answer to 2 questions:

1. What is, morally speaking, an AI capable of doing?

Agent→ being with capacity to act→ ‘agency’ is the exercise or manifestation of this capacity→ moral agency is an individual's ability to make moral choices based on some notion of right and wrong and to be held accountable for these actions. (Wikipedia)

Moral agents perform a voluntary behaviour, this means that we can explain the actions by analysing their mental states that produced such action. Involuntary behaviour on the other hand doesn’t produce actions and we have to find the reasons of such behaviours in what caused them.

Conditions for moral agency:

- Agent with internal state (mental state such as desires and beliefs) - Outward embodied event the agent does (exa he moves his body), which is caused by his internal states

- Outward behaviour causes an outward effect (on someone else)

2. How should we treat an AI? Patiency, as it was a patient and not an agent

We have to create moral machines so that they can make ethical decisions themselves. → → **Moor** identifies 4 categories of ethical agents:

1. Ethical-impact agents→ any machine that can be evaluated for its ethical consequences exa. A self-driving car

2. Implicit ethical agents→ machines designed in order to not have negative ethical consequences

3. Explicit ethical agents→ machines that reason about ethics using ethical categories as part of their internal programming, this is done by using various forms of deontic logic that have been developed for representing duties and obligations.

4. Full ethical agents→ they can make explicit moral judgments and are competent in justifying such decisions. This level of performance is often presumed to require capacity for consciousness, intentionality, and free will.

For Moor 3. Is the goal for machine ethics

Wallach and Allen identify two kid of morality:

1. Functional morality→ capacity of machines for assessing and responding to moral challenges. Technology with autonomy and sensitivity. They ought to have decision-making abilities.

Functionalist approach has critical difficulties:

- Testing: “we will need some metric by which to evaluate whether or not a particular device is capable of making the appropriate moral decisions in a particular situation”. Moral Turing Test?

- Anthropocentrism

- Slave ethics

2. Operational morality→ technology without autonomy and sensitivity

To produce the type of machine capable of **decision-making**, we can apply different ethical approaches and different ethical theories in the *process of design*.

1. Top-down approach→ having a set of rules that can be turned into an algorithm. Utilitarianism and deontological.

2. Bottom-up approach→ create machines that mimic a child’s moral development 3. Mixture of bottom-up and top-down→ (**Coeckelbergh**) proposes to build quasi moral robots that produce the appearance of being moral, they should imitate subjectivity and consciousness in a convincing way, that is what humans do. Humans, in fact, don’t have the proof that another person has mental states and consciousness, they just interpret the others’ behaviours as emotions.

Relational approach tells us that moral standing isn’t an objective property that we define just by judging the behaviour of others, we also need

discussion - in a human language thinking way- to interpret in a more complete way moral states. This means that robots might appear in different ways to different people in different situations, hece ethical attention passes from ontology, that deals with properties of an entity, to epistemology, that deals with our knowledge about the entity. It also passes from object to subject and from what things really are to ho we look at them.

For **Anderson and Anderson:**

We can give machines a human kind of morality, at first we might believe it will give us only advantages since they don’t have emotions and they might be more rational but in reality it has its limitations: moral rules conflict between them and emotions are required to make a complete moral judgement.

For **Johnson:**

Machines can’t be moral agents at all because they don’t have mental states, emotions or free will but they are not neutral like technical artefacts since they have autonomy and a kind of intentionality.

*Natural entities are different from human made entities:*

The first ones don’t have intentionality, they only have functionality, while the seconds have also intentionality, of course designed by humans who created them.

It is important to make such a distinction because we need to regulate human made objects with norms in order to prevent our future.

*Artefacts are different from technology:*

The first ones aren’t based on scientific knowledge while the others are, moreover the latter can be both tangible and physical objects.

How to think about Computers?

Computers might be thought to have all the conditions needed to have moral agency but it isn’t true; machines, that are human made entities, don’t have the human intentions to act, they have of course intentionality designed by who created them but this isn’t sufficient to make them morally responsible of moral appraisal.

NB! Computer systems are not moral agents, but they are a part of the moral world. They are part of the moral world not just because of their effects, but because of what they are and do. Moreover computers are closer to being moral agents than are natural objects.

NB! While talking about intentionality we need to clarify that systems have intentionality. They are poised to behave in certain ways, given certain input. The intentionality of computer systems and other artefacts is connected to two other forms of intentionality, the intentionality of the designer and the intentionality of the user.

The intentionality of computer systems is inert or latent without the intentionality of users. Users provide input to the computer system and in so doing they use their intentionality to activate the intentionality of the system.

The point of Johnson’s analysis of the intentionality of computer systems is twofold: - it emphasises the *dependence of computer system behaviour on human behaviour*, and especially the intentionality of human behaviour. While computer behaviour is often independent in time and place from the designers and users of the computer system, computer systems are always human-made and their efficacy is always created and deployed by the intentionality of human beings

- pointing in an entirely different direction, since computer systems have built-in intentionality, once deployed they *can behave independently* and without human intervention.

*Mindless morality:* (**Floridi**)

We could make moral agents dependent on :

- having a sufficient level of:

1. interactivity (response to the stimulus by change of state)

2. autonomy (ability to change state without stimulus)

3. adaptivity (ability to change the ‘transition rules’ by which state is change) - being capable of moral qualifiable actions.

An agent is morally good if its actions all respect that threshold and it is morally evil if some action violates it.

**Sullin**

pov on the three guidelines of moral agenthood: *full morality*.

- Autonomy→ if the machine is not under the direct control of any agent or user we say that it has a practical independent agency. The more autonomy they have, the more agency we attribute to them, moreover when such agency causes good or harm we can say the robot has moral agency.

- Intentionality→ we can’t prove the robot has a real intentionality in the strongest sense we think *but* If the interaction of the robot’s and environment causes the machine to act in a way that is morally harmful or beneficial, and the actions are calculated, then the machine is a moral agent.

- Responsibility→ whenever a robot behaves responsible to other moral agents, it has the duty to care for its patients.

Conclusions: Robots are moral agents when there is a reasonable level of abstraction under which we must grant that the machine has autonomous intentions and responsibilities. *The robots of today can be seen to be moral agents of a sort under certain, but not all, levels of abstraction and are deserving of moral consideration.*