Moral control and ownership in AI systems

Raúl González Fabre
Javier Camacho
Pedro Tejedor Escobar

AI systems are bringing an augmentation of human capabilities to shape the world. They may also drag a replacement of human conscience in large chunks of life. AI systems can be designed to leave moral control in human hands, to obstruct or diminish that moral control, or even to prevent it, replacing human morality with pre-packaged or developed 'solutions' by the 'intelligent' machine itself.

Artificial Intelligent systems (AIS) are increasingly being used in multiple applications and receiving more attention from public and private organizations. Practitioners must ensure that AIS are governable, open, transparent and understandable, able to work efficiently with people and consistent with human values and aspirations. The purpose of this article is to offer a fundamental mapping of the technological architectures that support AIS, under the specific focus of moral agency.

Through a literature research and reflection process, we explore questions related with four aspects of AI systems: Problem definition in the use of AI for complex issues (part 1); a review and taxonomy of artificial “reasoning” architectures (part 2); Influence of the data input and the data quality (part 3); AI systems in decision support and decision making scenarios (part 4). Finally, we offer some conclusions regarding the potential loss of moral control by humans due to AI.

This article contributes to the field of Ethics and Artificial Intelligence by providing a discussion for developers and researchers in order to understand how and under what circumstances the 'human subject' may, totally or partially, lose moral control and ownership over AI technologies.

Keywords: Artificial Intelligence, Moral Agency, Data Bias, Machine Learning, Autonomous Systems, Decision Support

LIST OF ABREVIATIONS

AAN: Artificial Neural Networks

AI: Artificial Intelligence

AIS: Artificial Intelligence Systems

AS: Autonomous Systems

DSS: Decision Support Systems

GAN: Generative Adversarial Neworks

ML: Machine Learning

RL: Reinforcement Learning

SAS: Semi-autonomous Systems

Introduction

The aim of the present article is to offer a systematic approach to understand how Artificial Intelligence is affecting our moral agency. True enough, the so called Artificial Intelligent Systems (AIS) can adopt options that, should a human take those, would be objects of moral study.

For that purpose, we divide our exploration in five main parts: first, we pose the problem in more detail; second, we introduce the ways of 'reasoning' of an AIS; third, as this 'reasoning' depends on data the AIS receives or captures, we discuss the issue of data quality; fourth, we present the ways an AIS may take part in decision-making processes leading to actions; and finally, we discuss some conclusions about moral ownership in schemes that include an AIS.

I. Practical and moral agency

1.1. Practical agency

We shall use here the BDI model of practical agency of an agent proposed by Michael Bratman (1987). This model has the advantage of being of a logical nature rather than a psychological or anthropological one. For that reason, it has been used to model decision making in AI-endowed agents (see Meyer & alia 2015), at the same time that it can also be used to understand human decision-making--actually it comes from what Bratman (1987) calls "a commonsense psychological framework" .

The model includes three elements:

- Beliefs of the agent: her representations of herself, other agents and the general environment.
- Desires[1]: states of the world--including herself--that are wished by the agent.
- Intentions: future actions the agent is committed to, in order to realize her desires given her beliefs.

From a logical point of view, ‘beliefs’ may be true or false, depending on the relation between their content and the aspects of the world they intends to represent. ‘Desires’ have the opposite “direction of fit” (Tiberius 2015:48): They do not intend to represent actual facts but to modify future facts, as to adjust them to what is desired.

[1] We keep here the word 'Desires' because it was used by Bratman and it continues being used in the related literature. However, it has a psychological-Humean taste that does not seem necessary. Maybe better than 'Desires' we should speak of 'Purposes', with the same logical content and less psychological charge.

‘Possible’ or ‘impossible’ are predicates applicable to desires, but ‘true’ or ‘false’ are not. Finally, ‘intentions’ are commitments, thus a kind of desire: they intend to adjust the world to a mental desideratum. They are special desires, however, because of the commitment force involved, that transforms ‘wish’ into ‘will’, so to speak.

Intentions refer necessarily to the future. They are organized in 'plans' more or less precisely defined along time. The rationality of those plans requires them to be consistent with the agents’ beliefs and desires, and with her other intentions--in consequence with the beliefs and desires those intentions intend to realize in turn.

1.2. Moral agency

Initially thought for people, Bratman’s model can be applied to both AI-endowed agents and humans. The implementation of the different elements and their interaction is however very different from one to another--and also among artificial agents with different internal architectures.

Starting from the anthropology of Xavier Zubiri (1986), we shall call ‘moral agency’, or ‘morality’ for short, the specific realization of practical agency in persons.

Zubiri describes the person as a peculiar kind of reality, open to self-definition by her own choices. As a consequence of her way of perceiving and choosing, the human agent owns the action chosen (she has preferred it over other possibilities available to her choice), and vice-versa (she is the one who has chosen that precise action, thus defining morally herself in the fact, eventually modifying herself--the ancient theory of virtue as a habit built through exercise[2]).

The first meaning of moral appropriation (from person to act) constitutes the classical basis for 'moral responsibility': the agent is also author that can be called to respond of her choice by others. The idea is already found in Aristotle, EN, 1109b. Such moral responsibility may in turn become the ground for legal responsibility[3].

The second meaning of appropriation (from act to person) relates morality with the subject's constitution as a practical agent. Morality requires certain psychological processes, mapped diversely by different authors (for example, motivational, cognitive, self-regulatory, enumerates Tomasello 2018: 661), that generate some kind of self-definition through choices. In consequence, it is not only a matter of behaving in certain ways, but also of the internal makings that result in the agent choosing that behavior and the internal consequences of so doing. Those internal makings and

[2] See for example Faucher and Roques (2018).

[3] We shall not enter in the much discussed issue of moral and legal responsibilities of AI-endowed systems. A good summary of both issues can be found in Chinen (2019).

consequences belong typically to the human psyche, and so 'morality' is a trait of humanity.

When applied strictly to people, Bratman's scheme allows to understand our particular complexity as moral agents:

- Beliefs: People form their beliefs about the world, themselves and other agents, by accumulating memories from their direct or indirect experience, as far as their mind allows. Even for the person, it is not always easy to identify all the beliefs in play in a moral decision[4], much less from which experiences they were formed.
- Desires: In each decision, the person may try to reach purposes of several kinds, some related to fully external states of the world, some others regarding relations with other people, and some related to her own self-definition as a person. Since Plato (*Republic*, 436a), it is known that our psychosocial constitution endows us with several sources of desires, not necessarily congruent among themselves.
- Intentions: The logical coherence between beliefs, desires and intentions often seems to break in the case of people (not to speak of the coherence among different intentions). When such incoherence shows up, it can well be that we are not being rational agents, or that we are being rational but with beliefs and/or desires hidden even to ourselves.

Human morality understood as a concrete implementation of practical agency has important differences with implementations in AI-endowed machines. Machines have to be manufactured, and so fully specified at least at the moment of their manufacture. Later on, they may evolve according to rules also determined in their building. Concepts like 'self-consciousness', 'experience', 'free will', "responsibility"... that are often assigned to moral agents, make little sense regarding AI-endowed machines.

As a consequence, we can speak of 'AI-endowed machine Ethics' only in an analogous way. The discussion on ascribing certain predicates (good, bad, better, worse, obligated, allowed, forbidden, indifferent...) to alternatives of decision, makes proper sense only in the case of moral agents, because it is their internal constitution what makes those predicates meaningful.

Applied to other practical agents, able to make decisions but with different internal architectures, either we discuss morality by reference to some human involved--the owner, the user, the programmer, the ruler--; or we are making an implicit

[4] We use the word 'decision' here to group together what Bratman (1987) calls 'intentions' and 'plans'.

'personalization' of the AI-endowed machine, maybe useful for rhetorical purposes but prone to confusion.

In both cases, the discussion to assign moral predicates may result in conclusions about the implementation of the BDI scheme in the agents under scrutiny. Not first what they must choose, but how they must be built in order to choose according to certain criteria in certain circumstances (see Wallach & Allen 2009).

1.3. Problems and messes

Moral rationality is not only a matter of optimizing a certain goal. Choosing the next adequate goal is part of the moral question posed to an agent endowed with morality. This is not an obvious task.

Hester & Adams (2017) notice the difference between a 'mess' and a 'problem'. In problems, the 'owner' and the 'goal' of the issue are well-defined. Someone is intending to maximize the achievement of a certain goal, given the available resources. In these situations, choosing rationally is a matter of optimization.

A mess is a set of interconnected problems with different owners (stakeholders we could say). Interconnection implies that a solution given to any problem in the mess modifies the terms of other problems within the same mess. In consequence, the problems are not initially all well-defined. But neither is the mess itself. Its contours may change: new problems and new stakeholders may come into the scene as we approach a certain problem in the mess.

Problems call for solutions, messes for management. Human morality comprises both: when solving the problems owned by her, the agent is also influencing problems owned by others linked to hers in messes. The management of any 'mess' is necessarily a moral issue, that requires decisions by an agent endowed with morality, that is, a human person.

1.4. The basic question of the paper

As conceived up to mid-20th century, machines merely constituted useful extensions of human action, increasing its range of possibilities in the most varied situations. They didn't modify essentially the morality of processes. Human beliefs, intentions and desires happened in the field of possibilities now broadened by machines.

Machines with AI (either alone or connected in networks of any topology) pose a new challenge. They may not only extend the field of human morality with more possibilities but also eventually replace genuine moral action with some machine choice of operation.

The 'morality question' arises in the insertion of an AIS within human plans, when it replaces or conditions the operation of one or more moral characteristics of the human action, in a way *de facto* difficult or impossible to control by persons. If that happens, the 'behavior' of the AI machine 'escapes' the field of human morality.

That issue is previous to the 'moral question'. Before discussing the moral predicates that may be rationally ascribed to a certain operation of an AIS (whether it is good, bad, etc.), that operation must take place within the field of morality. If the AIS has replaced or conditioned morality to the degree that is no longer acting fully, the moral discussion loses its basic meaning.

In this article, we explore the possibility of AI machines operating in ways that escape or condition heavily human morality. Our main focus is on the basic architecture of some modalities of AI, not in the particular uses of those modalities within specific AIS.

II. Artificial 'reasoning'

2.1. Rational Agents and their Universes

The paradigm that best suits a technical discussion about this new reality is the Artificial Intelligent System as a Rational Agent (Russell, Norvig & Davis 2010). With roots in Aristotle’s *Nicomachean Ethics*, we can model the system as an agent that assumes its environment, and based on that, adopts practical decisions followed by actions, in what we could name after “practical reasoning” (McCarthy 1958)[5].

A rational agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators. The environment is all the agent knows external to itself; so, it may be safely called the agent’s Universe. From a mathematical point of view, the agent’s behavior is a function that maps the perceived sequence from Universe to the sequent action in the Universe.

[5] This author was the first in adopting this terminology. It may also be the first paper to propose common sense reasoning ability as the key to AI

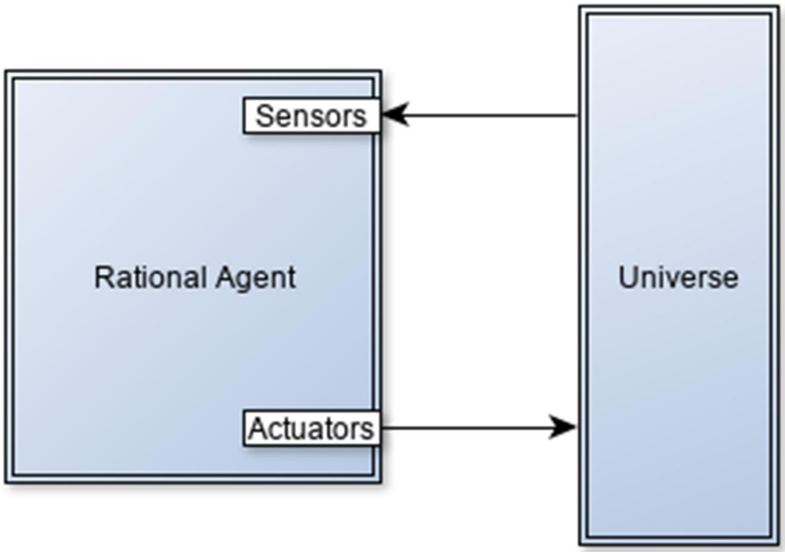

Figure 1 Agent-Universe interaction

An agent using a rigid table mapping from inputs to outputs could hardly be considered as “intelligent”. Sensors are never 100% accurate, and the corresponding universes cannot be generally taken as fixed. So, a more complex definition is necessary for a higher level of intelligence of the phenomenon.

One of the first steps involved considering levels of certainty, like in MYCIN (Shortliffe 1977). Other approaches were provided, and one that carried a lot of consensus was that Rational Agents should be able to learn from the Universe. The increasing capacity in data storage and computer power increased greatly the ability to learn from examples.

A closer view to the Universe surrounding the rational agents would reveal a moving environment with heterogeneous sources of variation, as seen in Figure 2:

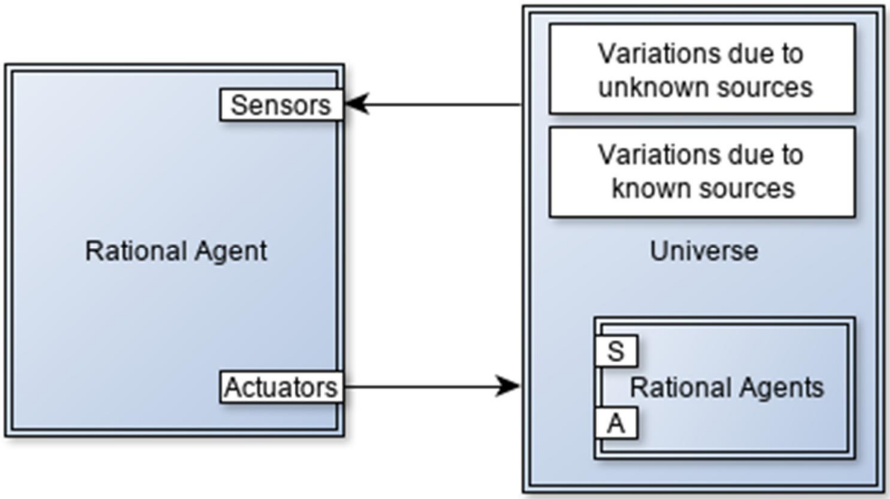
Variations due to unknown sources
Variations due to known sources
Universe
Sensors
Rational Agent
Actuators
S
Rational Agents
A

Figure 2 Relationship between Agent and Universe

The Universe that sends signals to the Agent and that receives its responses, is most usually a *moving* environment. We can find four types of sources of variation:

- Variations due to transient conditions of known sources. “Known” here means that we know the origin of the variation, and we know its behaviour, because it is produced by a phenomenon that follows some known rule (e.g. alternate current is known to change sign a number of times each second), or because we control that source of variation (e.g. a man switches on a light).
- Other sources that we may not know of, and we deem them as noise, hoping it will not bring bias to the input signals in any systematic way. This includes also any inaccuracy that the agent may experiment while acquiring the external signal with the sensors.
- Between type 1 and 2 there is another class of sources that may be predictable to some extent, but they are not determined, and carry some random variations with them e.g. the blow of the wind.
- In a growing number of scenarios, a rational agent operates in a Universe populated simultaneously by other rational agents (“secondary agents”). The behaviour of these secondary agents and the variation they bring to the Universe is different of those in case 1, 2 and 3, because each one is guided by a “utility function”, that expresses the value that agent aims to maximize from the Universe. In consequence, those secondary agents may react to the behaviour of the rational agent under study with interactions driving the actions of our rational agent to unintended results.

2.2. Expert knowledge

The relevant question to be answered in this paper regards the behavior of the agents. The simplest kind of behavior is just a list of condition-action rules, like “if you perceive this, then do that”. The first Expert Systems in history chose this type of approximation. They were used to reflect in a simple and transparent way, the knowledge of the experts in the subject matter (thus the name). A new career was devised, the “knowledge engineer”, to translate the knowledge of the expert into the function of the Rational Agent. The resulting Expert System aimed to make a “twin” of the way in which the mind of the expert modelled her domain, her “Universe” of expertise, and adopted decisions. The embodiment of this “digital twin” (Gelernter 1992) in a software system, acts as an internal model of the Universe for the Rational Agent. We will call it a “model”, that connects the data collected by the sensors to the actions.

Modelling an environment using a static model, like a long list of “condition-action” rules, fell short for any new variation not foreseen. The models so built couldn't cope with so many circumstances as the sources of variation could produce. Even if it could with all the past and present conditions, it might fail in the future. In order to get over this limitation of the initial Expert Systems, nowadays it is expected from the agent to be able to learn from new perceptions, at least to some extent.

2.3. Learning Agents

A Learning Agent mirrors the environment from the perceptions of its sensors. It needs the capacity to learn from examples. In fact, the ability to learn from an exponentially increasing quantity of data is what has fueled the growing of AIS.

This model captures the perception of the state of the Universe. After that, the Rational Agent needs a way to take a decision, driven by a goal or situation desirable to be achieved. It may be, for example, a place to arrive to for an autonomous car in a safe and quick manner.

In order to asses which is the best action to adopt, the Rational Agent is endowed with a “utility function”. This function is essentially an internalization of a performance measure. If the internal utility function and the external performance measure are in agreement, then an agent that chooses actions to maximize its utility will be rational according to the external performance measure. Figure 3 represents this separated internal function of the AIS.

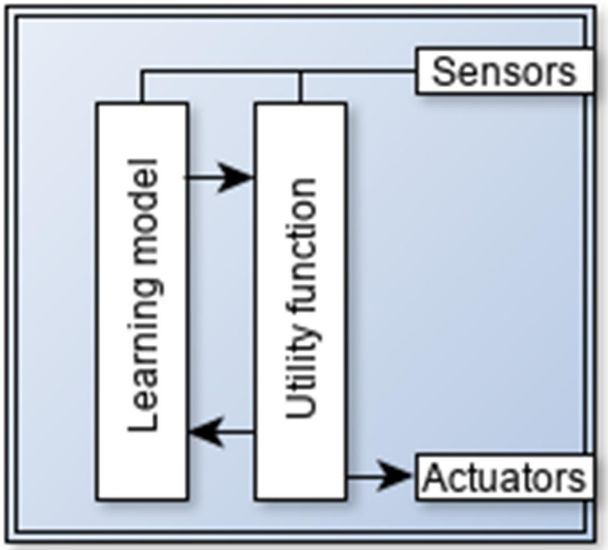

Figure 3 Agent with Model and Utility function

Utility functions are useful to manage some conflicting situations (Russell, Norvig & Davis 2010:72). First, when there are conflicting goals, only some of which can be achieved (for example, speed and safety), the utility function specifies the appropriate tradeoff. Second, when there are several goals that the agent can aim for, none of which can be achieved with certainty, utility provides a way in which the likelihood of success can be weighed against the importance of the goals. Partial observability and stochasticity are ubiquitous in the real world, and so is decision making under uncertainty. Technically speaking, a Rational Utility-based Agent chooses the action that maximizes the expected utility of the action outcomes—that is, the utility the agent expects to derive, on average, given the probabilities and utilities of each outcome.

What we have learnt about utility functions allows now to take another view to the Universe of the agent, populated with other agents with utility functions that may conflict with one another, so the task of the learning model and the utility function grows in difficulty. Some agents may be known, while others may be unknown, but influencing the Universe nevertheless. Utility functions may be built taking other agents' goals in account, in a collaborative or adverse way. This makes the assessment of the utility function much more complicated.

2.4. Machine Learning

The technical capacity that has led to the growth in the usage of AIS is that of learning from examples, not from experts. We may now take a more technical view to this capacity, to understand it better. It is called “Machine Learning”, as a generic name.

In Machine Learning there are different models that generally fall into three different categories: (1) Supervised Learning, (2) Unsupervised Learning and (3) Reinforcement Learning.

Supervised Learning: suppose you have a good quantity of past situations for which you know the expected output of the model. This output may be categorical (e.g.: vote to different parties), or numerical, either continuous or discrete. You can train a machine learning model by presenting the examples and adjusting the model to reduce the output error. You need a supervisor, more knowledgeable than the learning system, capable of tagging new examples to make new learning.

Two types of task fall under this kind of learning: classification and regression:

- Classification: if you have a set of observations, each one pertaining some category, the model must learn which of the characteristics drive the sample to be part of each category.
- Regression: the aim is to predict and forecast numeric response values given the values for each sample.

A very popular technology for this kind of learning is called Artificial Neural Networks (ANN). It is based on a conceptual model of a neuron and its workings. We will describe it here because of its importance, as the basis for the most popular AIS.

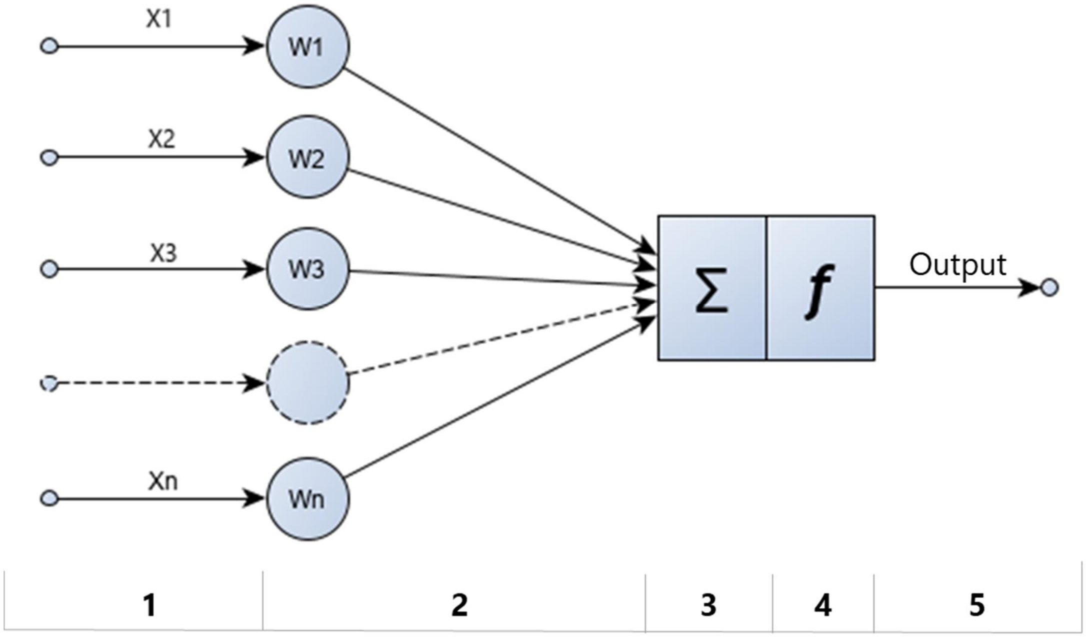

Figure 4 Schema of an Artificial Neuron

An Artificial Neuron (Figure 4) is the elemental brick to build the ANN. It has five main components. From left to right, these are:

1. Input values, each associated with an input node.
2. Weighted Connections. Each connection from the input node has a weight (w_i) associated that can be any real number. These are the values that are adjusted in the training of the artificial Neural Network.
3. Values from the input nodes times the weights ($x_i \cdot w_i$) are added.
4. This is the input for a transfer or activation function *f*, used as a filter (only operates if the sum of the inputs exceeds a threshold), or as a regressor.
5. The result is found in the output node, which may be the final output, or the input to the next layer of artificial neurons.

As for the training algorithm[6], we typically have a large training data set made of many input examples associated with their expected 'real' outputs as determined by a human operator. The model first makes a forward pass, using the initial weights, producing a prediction of the output to be compared with the expected output. Using this comparison, a loss function is calculated and used to update the weights. This phase is called “Back Propagation” and is used to fine tune the weights so that the loss decreases thru consecutive passes.After each training phase, a testing phase using

[6] For a deep discussion on ANN and the fundamentals of their learning, see Bishop, C. 1995.

another set of examples is run to see whether the ANN model works. Other strategies can be applied to continue tuning the model.

ANN is the base for the most popular Artificial Intelligence applications today, those of Artificial Vision, that use a special kind of ANN for deep learning in images, Natural Language Processing, that uses recurrent neural networks, and GAN (Generative Adversarial Networks) to generate images and sounds that are like real.

Unsupervised Learning: When data lacks a target categorization or there are not enough labelled examples, another type of techniques are used to find hidden patterns in the data. The most popular procedures under this category are clustering and association.

- Clustering: used to group together samples that are similar. A common useful measure for the goodness of the clustering is called “entropy”. The clustering techniques aim to minimize the internal entropy in clusters, while maximizing the entropy between clusters.
- Association: this task aims to find the rules that connect different samples. For example, when X happens usually also Y happens.

So much for the building of models based on cases. They are focused in modelling the Universe. Now we are going to take a brief look at a different approach, based on the learning by examples of the utility function.

Reinforcement Learning: the system is presented with a very vague set of rules that guide in the “how” of the action process, but not in the “what” should it choose to do. Then a set of situations is presented. From the decisions adopted, the system gets a “reward” or a “punishment”, and this way it learns how to behave to maximize the reward. Incipient as it is, Reinforcement Learning brings the possibility to build AIS that learn their utility function directly from the Universe they operate into.

2.5. Two problems for moral agency in the 'reasoning' of AIS

Moral agency requires some intelligence of the situation, in order to act conscientiously in it. The description already made of the internal 'reasoning' of an AIS poses two main problems:

Limits of application: A well-programmed AIS is reliable only within the limits of the Universe where it was trained. If faced with real-world situations that fall outside those limits in a relevant way, the AIS can produce unpredictable/undesirable outputs. In example driven AIS, as presented here, the limits of the Universe are defined exclusively by the samples the AIS is trained with, not by any predefined rule. If an AIS is to produce an adequate response, it must be trained with samples covering all situations where it can possibly operate.

Going out of limits can be understood in two ways: a more quantitative one, if the input values obtained from the real-world situation are out of the verified dominion of the Universe. This can be tricky, because the extreme limits of that Universe may be considered within its dominion, but not enough cases were available for the AIS to learn adequate responses. The financial crisis of 2008 exposed a clear example, when financial markets were modeled using Bell curves, adequate for small distances from the mean but grossly inadequate for extreme events, where very few cases for training AIS were available.

Going out of limits can be understood also in a qualitative way: AIS consider only the situation as far as they have sensors for. In the terms we presented in 1.2, they are able to cope with well-designed 'problems'. But those problems are often embedded in 'messes' where the problem makes sense. Therefore, if the 'messes' where the AIS is inserted change, the definition of the problem may also change: which input is relevant, which utility function would be really useful, and in consequence which output would be adequate. The AIS may result 'blind' to new aspects of the 'mess' not taken into account in its design.

These two modalities of the problem apply equally to Expert and to Learning Systems. In both cases, some human (moral) mind has to check whether the problem to be solved remains the same, and the AIS is working within its dominion of reliability.

Traceability (transparency): While expert systems are internally well known at least by their programmers (after all, they encode explicit expert knowledge), AIS based on machine learning may not be. Their relation input-output can be mapped within their training Universe, and it extends to anything in between. But how do they reach that functional map, is often impossible to understand even with full access to its internal workings, due to problems in legibility of ANNs for human minds.

This becomes more problematic for systems that use Reinforcement Learning. An ANN that operates under RL changes its weights and thresholds automatically; soon its workings may become opaque even for its programmer. Not even the actual map input-output is well known, because it is an internal result of the system itself, often without need of external assistance.

Taking moral decisions in systems that include AIS requires to know how the system will operate on every alternative under consideration. As the AIS grows less understandable to a human mind, it also makes the moral evaluation of alternatives more difficult to perform.

III. Data quality

3.1. Data as a key element of AI

Machine learning algorithms are devised to find a function to map the features of a given input data to a desired target output, for which in some cases they need a large amount of input data for “training” or “learning” purposes. These algorithms “learn” from the data provided, and therefore their Universe is built upon this data, no less no more. This means that data themselves play an important role in the definition of the capabilities of any such algorithm. As a matter fact the major advances in AI have been possible due to the availability of vast amounts of digital data, and therefore “Without data, there is no AI” (Bowles 2018: 62).

Data sets are indeed a key part of AI systems, and as such they are receiving increased attention from researchers and companies (Vakkuri 2018). Both supervised and unsupervised learning depend on training data, i.e., known datasets for practicing and tuning the machine-learning model (Broussard 2018). Even in the case of reinforced machine learning, the algorithm finds reproducible patterns in the data, so if the data are distorted or skewed, then that is the pattern that the algorithm will learn (McQuillan 2018). The increase in use of machine learning algorithms has brought an even higher increase in the need for data, “Big Data", which implies an abstraction and disarticulation of data about individuals whose activity in the digital space is the source of these data (Markham 2018). Furthermore, in order to be usable, data must be treated and conditioned with tools, and therefore the “rawness” of data may be disrupted in various manners (Ekbia 2015), which might introduce further distortion in the data sets.

For the purpose of this article, we have classified the issues of data according to the following criteria: origin, processing, aggregation, post-processing of data sets and cumulative effects of those (see Figure 5):

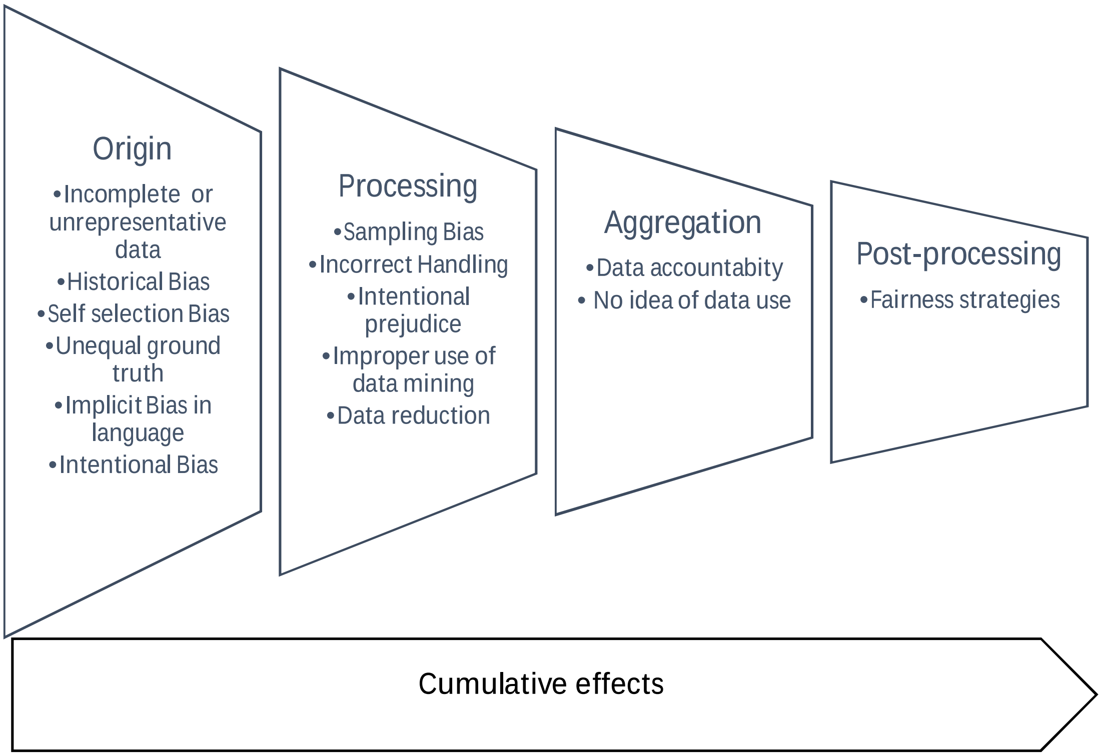

Figure 5. Classification of potential issues related to data sets

3.2. Origin of Data

Every company, in a declared or undeclared manner, benefits from capturing more data. Machine learning is only effective when the training data sets are large enough, so both the ethical question of consent and the reflection about the quality of the data sets might be marginalized by this need of scale (McQuillan 2018), left apart the moral consideration about respect for human dignity, which requires that all people are treated with respect due to them as individuals, rather than merely as data subjects. In addition to the ethical reflection about the obtention and use of data, data sets might be incomplete or unrepresentative, because they have not been properly selected: “it is an error to assume ‘people’ and ‘twitter users’ are synonymous” (Boyd & Crawford 2012:668).

Data might be tainted with different kind of bias, understood as a prejudice for or against something or somebody, that may result in unfair decisions (AI HLEG 2018). There are different kind of bias, such as historical bias or self-selection bias, which occurs when we compare people who made different choices without considering why they made these choices. Data might also simply represent an unequal ground truth, i.e., a non-biased reality in which capacities or risks are unevenly distributed between different groups (Hacker 2018). Another effect is the implicit bias in language (Caliskan et al. 2017), which is to be considered when utilizing word embeddings or semantics representation of words in AI systems.

There might be also cases of intentional bias, that could be the result of using as training or test data, malicious data that have been intentionally fed into the system, for example in the case of chatbots training affected by trolls (Bowles 2018).

3.3. Processing of Data

Extracting data is not a passive act. It implies multiple choices of which data to collect, what to omit, how to process them and each of these actions has its own implicit assumptions (Bowles 2018). Preprocessing of data might include data integration, data cleaning, data transformation, data reduction and discretization (Mnich 2018). Several issues might arise at this pre-processing stage: sampling bias, causing that part of the population is misrepresented; incorrect handling of training data, caused by incorrect labelling (implicit bias in human coders) (Hacker 2018); intentional prejudice, when intentionally trying to achieve unfair, discriminatory, or biased outcomes in order to exclude certain groups of persons, by explicit manipulation of the data.

There is another issue related with the pre-processing of data related with an improper use of data mining techniques, that is analyzing data without a clear direction. Data-mining algorithms are programmed to look for trends, correlations, and other patterns in data, which may cause to invent theories or group data with criteria that have no ground reason behind: “We think that patterns are unusual and therefore meaningful. In Big Data, patterns are inevitable and therefore meaningless” (Smith 2018: 80).

3.4. Aggregation of Data

'Big data' is not simply data massively scaled-up in quantity, but rather datasets connected through algorithmic analysis, forging unpredictable relationships between data collected at different times and places and for different purposes (Metcalf et al. 2019). Aggregation changes the data landscape (Bowles 2018). If full AI accountability implies accounting for the origins, construction and use of training and test data (AI Now 2018), then data aggregation cannot be underrated, and some specific techniques for complete data traceability should be in place.

In relation to this, there is a risk of non-fully considering the potential data use, based on the deceptive premise that knowledge derived from data analytics is true (without further assessment) because the objective qualities of statistics and the size of the data set, which could imply that decision making and judgement are removed from the equation (Markham 2018). This might be specially challenging when using data output as data input in the "mess" architecture described above.

3.5. Post-Processing of Data

To counter some of the above-mentioned issues, efforts have been done towards the development of fairness-preserving algorithms, which seek to provide methods under which the predicted outcome of a classifier operating on data is fair or non-discriminatory for people based on any "sensitive" attribute (Friedler et al. 2018). The goal is to diagnose and mitigate bias applying methods such as anti-classification (the model does not depend on sensitive attributes in the dataset), classification parity (predictive performance of the model is equal across groups that are defined by sensitive attributes), and calibration strategies (ensures that outcomes do not depend on sensitive attributes) (AI Now 2018). However, these techniques need to find a suitable trade-off between accuracy and fairness since, in some cases, reduction of bias will also imply a decrease in prediction accuracy. Besides, fairness-aware algorithms tend to deliver different outcomes depending on fluctuations in dataset compositions, implying that post-processing fairness interventions might be more brittle than previously thought (Friedler et al. 2018).

3.6. Additional Effects of Data Input for AI systems

To add another degree of complexity, the effects described above might be cumulative. For example, we can imagine that a certain post-processing fairness strategy might be applied to an aggregated dataset, coming from several datasets, one with sampling bias error on top of a set of historically biased data, another incomplete and another with incorrect handling, such as improper labelling. In those cases, the potential biases propagate with unforeseeable consequences and loss of control over the moral decision.

In order to fight data bias, AI should always be applied transparently, in order to understand, monitor and suggest improvements to algorithms; it is also suggested to include diversity among AI developers, in order to address insensitive or under-informed training of machine learning algorithms, and to foster collaboration between engineers and domain experts who are knowledgeable about historical inequalities (Caliskan et al. 2017).

Every kind of bias has a different solution and therefore the integrity of the data gathering needs to be ensured. Even when removing some types of bias at data collection, the identification of the bias has to be documented and the original data must be kept in record (AI HLEG 2018). Data traceability for the various inputs (training and test sets), additional testing for "fairness forensics" and more active intervention is needed in order to minimize potential undesired effects coming from data input.

The broad AI community is now well aware of problems of fairness, bias and discrimination as a result of data input, as it is shown by the number of initiatives on the topic: Fairness 360 by IBM, What-if tool by Google, fairlearn.py by Microsoft, Fairness-flow by Facebook, etc. (AI Now 2018).

Yet, there are a number of unsolved concerns about how to address this issue: which is the right way to de-bias an AI system? Should bias always be eliminated? Under which circumstances? Who is to make the implicit assumptions about what is and is not fair, in order to apply the proper fairness strategy to each situation? Furthermore, the proliferation of observational fairness methods through algorithmic treatment, which are not completely stable, as explained above, might provide a sense of false security (AI Now 2018).

IV. Decision making

4.1. Information and action

AIS are trained and fed with data (part III) and operate on them 'intelligently' (part II) in order to help or produce an output valuable to its user. The output of AIS may include both information and action:

- Information as input to another operator, human or artificial, that will process it to produce further information for third agent(s) and/or an action executed by itself.
- Action consisting of physical action--for example in a robot-- and/or decisions made in informational networks--for example, trade in an electronic market, assignation of rights in an institutional context, etc.

The decision-making process is practical in nature; it ends when an action is produced. The decision-maker is usually an operational unit--human or machine--, but the provision of information for its decision making may be quite complex in design, including other operational units networked in different ways.

4.2. Machines and human decision-makers

AIS and people are different as decision-makers. Differences often mentioned are:

- AIS are far more able than people to perform calculations. They can consider more information, process it quicker and execute decisions (being the case) almost instantaneously. They can act on patterns that people wouldn't notice and, on the opposite, discard as statistically insignificant or overfitting, patterns that people believe to see in the data.
- AIS have a different build from people. Computers are inorganic and unemotional, while people are organisms endowed with emotions. As a result, computers are much more regular than people as decision makers. For the same initial setup and information history, they will produce always the same output.
- AIS are less flexible than people. They can only calculate on the input they are programmed to consider, while people are able to make decisions based on their full life history, including their personal experience, social interactions, theoretical --even conflicting--backgrounds, etc.

And, most important for our purpose:

- People are moral agents properly, while AIS can be called 'moral' only analogically (part I).

AIS are being used in three different ways with regard to the decision making that leads to an action:

- Decision Support Systems (DSS): the AIS offers a human agent processed information and even suggestions about the decision to be taken. The final decision-maker is human.
- Semi-autonomous Systems (SAS): controlled most of the time by an AIS[7], which is regularly the final decision-maker. In some situations, however, identified either by the AIS or by a human operator, the control changes to the latter, who then becomes the final decision-maker. Usually those are situations deemed too complex to rely solely on the AI for the decision.
- Autonomous Systems (AS): controlled always by an AIS, that makes all decisions in all circumstances.

All these systems have common problems related to information, recounted in part III. If the data received is faulty for some reason, we can expect the system's decision, or intervention in a human decision, to be also problematic. If the processing of that data is opaque (part II), the decision will also be, in the sense that it will be impossible to

[7] The contrary is also frequent: A machine under the regular control by a human operator, that passes to an automatic system in case of catastrophic failure of that human operator (for example: in case that the driver of a car becomes distracted or asleep and the car threatens to leave the road). This are rarely AI systems: they don't have time/experience enough to 'learn' from their own performance. They are rather emergency, fully programmed mechanisms.

trace back how it resulted from the data. Here we are going to leave aside those questions, already presented, and discuss the aspects related to decision making itself, from the point of view of moral agency.

Regarding the possible loss of moral agency in the final decision-making step, we find problems of two kinds:

- The AIS often implies a silent choice about how to tackle a certain 'mess' (see 1.2): which problem is to be sorted out next and on which informational basis. In consequence, the AIS carries with it a certain framing of the problem.
- Additionally, to the extent that some autonomy is granted to the system (null in DSS, full in AS), the procedure for finding a solution to the problem may also be discharged from moral agency.

4.3. Decision Support Systems (DSS)

Having more reliable information to make a decision should improve it, as Bayesian statistics shows (Silver 2012). DSS are designed to provide such information, leaving the final decision to a human anyway.

This is the least problematic use of AI machines for decision making. The human decision-maker needs not follow the suggestion made by the AI machine, if any, nor consider the information provided by the machine as the clue for her decision--we could call it an implicit suggestion to the human decision-maker. She can reframe the problem or reintegrate it in a certain 'mess', consider additional information not provided by the DSS, exercise her moral judgement, utilize her own heuristics...

The problems related to moral agency are thus not essential to DSS. All of them emanate from a renunciation by the human decision-maker to exercise her moral agency in front of the problem, renunciation 'helped' maybe by the DSS.

This is not a new thing: blindly taking *prêt-a-porter* solutions to decision problems is a usual human heuristics. Those solutions can be provided by a behavioral code with 'if-then' clausulae (Boddington 2017) , by routine or habit (Kahneman 2011), by a figure of authority (Gibson 2019), etc.

Using a DSS easies the renunciation to moral discernment in some concurrent ways:

- Simplification: The DSS information can be understood as the most relevant, even the only relevant input for the decision. Incorporating additional, different information requires not to accept that implicit simplification.

- Speed: if the decision has to be made quickly, the human decision-maker may find it handy to simply accept what is implicitly or explicitly suggested by the DSS. Exercising complex moral judgement may require time and effort.
- Justification: it is often easier to justify a decision in front of others (supervisors, for example) if it was suggested or supported with data from the DSS. The DSS proposal acts then as a 'default'. If the human decision-maker separates herself from that 'default', she has to justify it; while following the 'default' rarely needs additional justification.
- Authority: when separating from a DSS explicit or implicit proposal, the human decision-maker is somehow challenging the authority behind that DSS. This may not only be the 'expert knowledge' of the programmer but also, and more important, the institutional authority that adopted that precise DSS.

4.4. Semi-autonomous Systems (SAS)

The SAS operate autonomously in many situations, but are able to pass the operational control to a human under certain circumstances detected either by the SAS or by the human operator. Zilberstein (2015) differentiates two kinds of SAS:

- SAS-I are systems where the human actions are not factored in the algorithmic design of the system. They pass control on to humans in certain circumstances and, when retaking it, simply start 'anew' from the situation the human operator has defined with her actions.
- SAS-2 are systems where human actions have been factored into the algorithmic design of the system. For example, as predefined branches among which the human operator has to choose.

A SAS makes two kinds of decisions:

- on the one hand it acts autonomously when it has the control, in the way it was programmed to--including self-modifying through autonomous reinforcement learning, if it is the case;
- on the other hand, in certain situations it passes the control on to a human operator and/or takes it back from her. When transferring control, it often gives information to the human operator for her to go on deciding about the operation.

When the SAS has the control, the moral agency problems are similar to the ones of an autonomous system (AS). When the SAS has passed the control on to the human operator, we may find moral agency questions similar to a DSS, if it has provided her with some decision-oriented information.

The specific moral agency problems for a SAS are found in the transfer of control back and forth between the human operator and the machine. In case that the AIS is receiving the control, the situation--the Universe as perceived by the sensors in the system--must be such that the AIS can perform well in it. If the system is a SAS-I, this cannot be ensured by the system itself, which will try to work in whatever the conditions it is put in. Reliability has thus to be guaranteed by the human operator.

In the case that the human operator is receiving the control, the problem of moral agency consists in her ability to assume it. That depends on the speed of the transfer, her physical and mental state at the moment, the information provided by the SAS and her (learned) capacity to use it and operate the machine...

That is not different from what already happens with many non-AI endowed machines, that operate according to a fixed automated program but transfer the control to a human supervisor in case that any parameter goes out of predefined intervals. Also, in those cases the human operator, used to a routine of automatic working of the machine, may not be ready for undertaking control at the necessary moment. An AI-endowed machine may be different only in that, thanks to its learning process, it may keep the system more often within the boundaries where there is no need for transferring control to a human operator. This may then get even more used to 'doing nothing' and less attentive to receiving the control.

Additionally, it must be programmed what happens in case that a SAS tries to transfer control to a human operator but, after some time, this operator has not taken it. The system may then either get stuck or operate fully as an AS.

For the rest, the presence of a 'human-in-the loop', definitory of SAS, allows for taking advantage of the complementarities between the differential characteristics of machine and human decision-makers (see 4.2 above). SAS may avoid (or tackle) mistakes that a human operator is more prone to do because of its limited memory and computing ability, physical constitution and emotional nature; while the human decision-maker may add her flexibility, professional experience and general ability to place the situation in a broader 'mess' context.

4.5. Autonomous Systems (AS)

By definition, autonomous systems need not a human operator. Their decision-making processes are fully preprogrammed and dependent on the information they receive via their sensors. They can 'learn' and thus change the basis for their decisions, even the algorithms for decision-making.

There is an obvious problem of moral agency here. Unless all possible information sets--along time--and internal configurations have been considered, in which case the programmer is doing all the moral decisions beforehand, the AS may operate in an eventually unpredictable--undecipherable--way.

The problem of traceability (transparency) described in 2.5, becomes more serious for AS whose output is an action. Leaving the human operator 'out-of-the-loop' implies in many cases losing moral control over the AS. Only when negative consequences are detected, the human operator may take control of the system or unplug it and gain the control that possible over the functions the AIS was in charge of.

Where those negative consequences may be catastrophic, there is a justified suspicion related to full AS. Only if no fatal consequences may be expected, AS can be trusted with a certain function. That's one reason, among others, why SAS are generally considered preferable to AS.

4.6. Information integration

Salgues (2018:10) observes: "the notion of information integration into processes and actions is more important than artificial intelligence, as we know it these days." This applies perfectly to the decision systems we have just mentioned. Devices that realize or support some kind of action--either physical or purely decisional--, are as strong as the weaker link that provides them with information. Networks with redundancy built into them are at most as strong as the strongest redundant mechanism in their weaker link.

Sometimes the weaker or the least reliable link in a decision chain or network is the human-in-the-loop. But, as mentioned above, it may well happen that the human operator precisely contributes experience and placement of the problem within a 'mess', far beyond what an AIS can do. In the case of redundancy mechanisms, her role may be essential to ensure that the best decision is made when diverse 'branches' of that redundancy mechanism give different recommendations.

V. Conclusions

In this article we have offered a fundamental mapping of the technological architectures that support AIS, under the specific focus of moral agency. Designed to assist, guide or directly adopt decisions, some AIS technologies present a potential risk of shifting the 'locus of control' from the human to the 'intelligent' machine. They replace or condition the operation of one or more moral elements of the human action, in a way *de facto* difficult or impossible to control, even to know, by persons.

To summarize our findings, we present them according to the essential elements of human morality that may be affected (see 1.1):

5.1. Beliefs

AIS collect and process the information relevant for maximizing their utility functions, as defined in their logical architecture. Their 'learning process' may be designed as Supervised, Unsupervised, or based on Reinforcement while on-the-go (see 2.4).

In all three cases, it depends on the Universe of cases the AIS is fed with. Any problem with that set of cases is automatically incorporated into the AIS operation, in a way easily unknown to the user of the system. We have mentioned problems related to:

- the origin of the data: consent, quality, completeness, representativeness, different kinds of bias implicit in the data themselves (see 3.2);
- the processing of data: sampling bias, incorrect labelling, intentional manipulation, improper use of data mining techniques (see 3.3);
- the aggregation of data from different origins in time, place, procedure of recollection... each one blindly involving its own options (see 3.4).
- the postprocessing of data with bias-cleaning algoritms, which introduce another layer of moral criteria often unknown to the user (see 3.5 and 3.6).

Even if the quality of the data used for the learning of an AIS is good, its operation may be inadequate or unpredictable when it happens out boundaries of the Universe defined by those data.

Finally if, as it often happens, the internal architecture of an AIS includes neural networks, it becomes intelligible to human minds, even knowing it in full detail. Moreover if there are neural networks continuously reprogramming themselves through Reinforcement Learning.

5.2. Desires

The first step in any use of AIS is choosing the problem to be solved. This requires managing the ‘messes’ where such problems may be embedded (see 1.2). As far as AIS are included in the decision-making process, they may bring implicit choices about how to tackle a certain 'mess': which problems are to be sorted out next, how are they to be framed, on which informational basis they will be solved.

Those definitions have to be considered within the broader scope of a decision chain or network. They require an intentional management of the 'mess' where the problem makes sense, which is a typical work of human moral agency.

Having at our disposal an AIS ‘solution’ fosters the temptation of assuming the definition and selection of problems presupposed by that precise ‘solution’. If our adoption of the AIS is too quick and unreflecting--if we were looking for nails provided that we have a hammer--, it could well happen that we define and solve problems adequate for ‘messes’ or in times different from the one we are intending to tackle. Silently, we would be assuming a way of choosing and defining the relevant problems along with the AIS.

Not only the selection of the problem, but also the choice of the AIS utility function (see 2.1) poses a major issue relating the 'desires' (purposes) we are trying to achieve. Those functions define both the relevant indicators and how they are to be mathematically combined to get a number the AIS will try and maximize. In consequence, when adopting an AIS, the user is assuming not only the vision of the world (the definition of the Rational Agent's Universe) but also the particular objectives the AIS incorporates in its design.

5.3. Intentions

Moral agency is only possible when there is a 'human-in-the-loop' of the AIS, that is, only in DSS and SAS. In totally Autonomous Systems (AS), no human operator is needed for the system to operate and thus to eventually produce undesirable consequences or make networks of other operators to produce them (see 4.5).

However, even in Decision Support Systems (DSS), excessive trust in the recommendations made by the AIS may replace moral decisions with the default suggested by the system. Circumstances of simplification, speed, justification and/or authority, facilitate the inhibition of morality where it would be possible to discern if a decision different from the one proposed by the DSS would be better (see 4.3).

In Semi-Automatic Systems (SAS) the question of moral agency extends to the point of the change of control from the AIS to a human operator and vice-versa. Who is to decide in each situation becomes an issue of whether the system remains under moral control or not (see 4.4).

5.4. Perspectives

Moral agency requires a special implementation of the BDI logic, that corresponds to human psyche and has its characteristic openness to the world and to itself (1.2). If some of its elements are somehow nullified by an AI-endowed practical agent, or if strong conditionings are placed on them in a manner that the humans involved cannot detect, the resulting operation falls out of morality. That changes the character of ethical discussions, because we would be mixing in them both moral agents in a proper sense with mere practical agents without morality. The borders between substantial and analogical moral operation would have to be well established for the discussion to remain logically sound.

The problem is bigger because AIS often are, and increasingly tend to be, not single machines but complex networks of machines--eventually very different--that feed information and decisions into each other and to human operators. The detailed traceability of input-process-output at each node of the network is essential for it to remain within the field of moral agency. This matter is receiving much attention lately, for at least two reasons:

- Moral agency is at the basis of our system of legal responsibility. As AIS in complex networks become more essential for the functioning of our societies, the preservation of moral agency through them acquires bigger relevance.
- It is also important for the commercial development of AIS itself. Social approval is unlikely to be obtained for entrusting important functions to complex systems under which no moral agency can be really identified.

Much remains to be done, not only in the field of the basic architecture of AIS we have summarized in this paper, but also with regard to the question of moral agency through complex networks that include AIS and human operators.

References

Adam, F. and P. Humphreys (2008). Encyclopedia of decision making and decision support technologies. Hershey, PA, Information Science Reference.

AI HLEG, High-Level Expert Group on Artificial Intelligence (2018). Draft Ethics Guidelines for Trustworthy AI. Retrieved from: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=57112.

AI NOW Report 2018, Artificial Intelligence Institute. New York.

Aristotle and R. Crisp (2000). Nicomachean ethics. Cambridge, U.K. ; New York, Cambridge University Press.

Bishop, C. (1995) Neural Networks for pattern recognition, Clarendon Press, Oxford.

Boddington, P. (2017). Towards a Code of Ethics for Artificial Intelligence, Springer.

Bowles, C. (2018). Future ethics. East Sussex, United Kingdom: NowNext Press.

Bratman, M. (1987). Intention, plans, and practical reason. Cambridge, Mass., Harvard University Press.

Broussard, M. (2018). Artificial unintelligence: How computers misunderstand the world. MIT Press.

Boyd, D. and Crawford, K. (2012). "Critical questions for Big Data", Information, Communication & Society, 15:5, 662-679, DOI: 10.1080/1369118X.2012.678878

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. “Semantics derived automatically from language corpora contain human-like biases.” Science 356.6334 (2017): 183-186. http://opus.bath.ac.uk/55288/

Chinen, M. (2019). Law and autonomous machines : the co-evolution of legal responsibility and technology. Cheltenham, UK, Edward Elgar Publishing.

Ekbia, H. , Mattioli, M. , Kouper, I. , Arave, G. , Ghazinejad, A. , Bowman, T. , Suri, V. R., Tsou, A. , Weingart, S. and Sugimoto, C. R. (2015). "Big Data, Bigger Dilemmas: A Critical Review". J Assn Inf Sci Tec, 66: 1523-1545. doi:10.1002/asi.23294

Faucher, N. and M. Roques (2018). The ontology, psychology and axiology of habits (habitus) in medieval philosophy. New York, NY, Springer Berlin Heidelberg.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2018). "A comparative study of fairness-enhancing interventions in machine learning". arXiv preprint arXiv:1802.04422.

Gelernter, David Hillel (1992). Mirror Worlds. Oxford University Press, USA.

Gibson, S. (2019). Arguing, obeying and defying: a rhetorical perspective on Stanley Milgram's obedience experiments. New York, Cambridge University Press.

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, 55(4), 1143-1185.

Hester, P. T. and K. M. Adams (2017). Systemic Decision-Making Fundamentals for Addressing Problems and Messes. Springer.

Kahneman, D. (2011). Thinking, fast and slow. New York, Farrar, Straus and Giroux.

Markham, A. N., Tiidenberg, K., & Herman, A. (2018). "Ethics as Methods: Doing Ethics in the Era of Big Data Research—Introduction". Social Media + Society. https://doi.org/10.1177/2056305118784502

McCarthy (1958). “Programs with Common Sense.” Proceedings of Teddington Conference on the Mechanization of Thought Processes.

McQuillan, D. (2018). "People’s councils for ethical machine learning". Social Media+ Society, 4(2), 2056305118768303.

Metcalf, Jacob, Emily F. Keller, and Danah Boyd (2019). “Perspectives on Big Data, Ethics, and Society.” Council for Big Data, Ethics, and Society. https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/.

Meyer, John-Jules Ch., Jan Broersen and Andreas Herzig, "BDI Logics", in Ditmarsch, H. v. (2015). Handbook of epistemic logic. [London], College Publications.

Mnich, M. (2018). "Big data algorithms beyond machine learning". KI-Künstliche Intelligenz, 32(1), 9-17.

Plato, et al. (2018). The Republic. Cambridge ; New York, Cambridge University Press.

Russell, S. J., Norvig, P., & Davis, E. (2010). Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Salgues, B. (2018). Society 5.0. Hoboken, NJ, Iste Ltd/John Wiley and Sons Inc.

Shortliffe, E. H. (1976). Computer-Based Medical Consultations: MYCIN. Elsevier/North-Holland.

Silver, N. (2012). The signal and the noise: why so many predictions fail--but some don't. New York, Penguin Press.

Smith, G. (2018). The AI delusion. Oxford University Press.

Tiberius, V. (2015). Moral psychology : a contemporary introduction. New York ; London, Routledge, Taylor & Francis Group.

Tomasello, M. (2018). "Precís of a natural history of human morality." Philosophical Psychology 31(5): 661-668.

Wallach, W. and C. Allen (2009). Moral machines : teaching robots right from wrong. Oxford ; New York, Oxford University Press.

Zilberstein, S. (2015). "Building Strong Semi-Autonomous Systems". Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

Zubiri, X. (1986), Sobre el hombre, Madrid, Alianza.