AI & SOCIETY (2021) 36:289–303
https://doi.org/10.1007/s00146-020-01020-z

OPEN FORUM

Moral control and ownership in AI systems

Raul Gonzalez Fabre[1] · Javier Camacho Ibáñez[1] · Pedro Tejedor Escobar[2]

Received: 21 October 2019 / Accepted: 10 July 2020 / Published online: 22 July 2020

Abstract
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 the public and private organisations. The purpose of this article is to offer a mapping of the technological architectures that support AIS, under the specific focus of the moral agency. Through a literature research and reflection process, the following areas are covered: a brief introduction and review of the literature on the topic of moral agency; an analysis using the BDI logic model (Bratman 1987); an elemental review of artificial 'reasoning' architectures in AIS; the influence of the data input and the data quality; AI systems' positioning in decision support and decision making scenarios; and finally, some conclusions are offered about regarding the potential loss of moral control by humans due to AIS. This article contributes to the field of Ethics and Artificial Intelligence by providing a discussion for developers and researchers to understand how and under what circumstances the 'human subject' may, totally or partially, lose moral control and ownership over AI technologies. The topic is relevant because AIS often are not single machines but complex networks of machines 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. Moral agency is then at the basis of our system of legal responsibility, and social approval is unlikely to be obtained for entrusting important functions to complex systems under which no moral agency can be identified.

Keywords Artificial Intelligence · Moral agency · Data bias · Machine learning · Autonomous systems · Decision support

Abbreviations

AAN	Artificial neural networks
AI	Artificial intelligence
AIS	Artificial intelligence Systems
AS	Autonomous systems
DSS	Decision support systems
GAN	Generative adversarial networks
ML	Machine learning
MTT	Moral turing test
RL	Reinforcement learning
SAS	Semi-autonomous systems

1 Introduction

The present article aims to offer a systematic approach to understand how Artificial Intelligence is affecting the human 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 into the following parts: first, we make a brief introduction and review of the literature on the topic of study; second, we pose the problem in more detail; third, we introduce the ways of 'reasoning' of an AIS; fourth, as this 'reasoning' depends on data the AIS receives or captures, we discuss the issue of data quality; fifth, we present the ways an AIS may take part in decision-making processes leading to actions; and

✉ Javier Camacho Ibáñez
jcamacho@comillas.edu

Raul Gonzalez Fabre
rgfabre@comillas.edu

Pedro Tejedor Escobar
pedro.tejedor@gmail.com

[1] Universidad Pontificia Comillas, Madrid, Spain

[2] Instituto de Ingeniería del Conocimiento, Madrid, Spain

finally, we discuss some conclusions about moral ownership in schemes that include an AIS.

1.1 Antecedents

The research field of Ethics in Artificial Intelligence (AI) has attracted much interest recently (Jobin et al. 2019). there is a growing consensus that digital technologies are legitimate objects of ethical concern (Greene et al. 2019), moving away from the technological neutrality view of the last decade. There are broadly two areas of research: AI Ethics and Machine Ethics (Winfield et al. 2019).

AI Ethics is mainly concerned with the governance of these systems and focused on the psychological, social and legal aspects of the challenges they pose (Yu et al. 2018). Such ethical governance pursues the development of a set of processes, procedures, cultures and values to ensure the highest standards of behaviour for both individual designers and the organisations in which they work (Winfield and Jirotka 2018). This branch is concerned with the ethical application of AI systems and has already led to the development of ethical principles and sets of guidelines (Jobin 2019).

Machine ethics (also referred sometimes as Ethical AI) is more concerned with the questions of whether and how AI systems can behave ethically (Winfield et al. 2019). The research in this field spans both philosophy and engineering. It explores decision-making mechanisms, in individual or multi-agent environments, how to represent ethical references by agents, or human–machine interaction (Yu et al. 2018). This paper is framed within this field of Machine Ethics.

AI systems are increasingly supporting human decision making, or they make decisions autonomously (Rossi and Mattei 2019). There is an urgent and real need for a functional system of ethical reasoning as AI systems are ready to be deployed at a massive scale (Charisi et al. 2017). There remain, however, two main challenges for that development: defining and formalising the ethical issue (philosophic) and implementing some degree of moral reasoning in autonomous systems (engineering) (Winfield and Jirotka 2018). The purpose of this article is to explore some fundamental ideas connecting these two challenges ahead.

1.2 A brief review of the literature on moral agents

Moor (2006) established a distinction between implicit ethical agents, that is machines designed to avoid unethical outcomes, and explicit ethical agents, that is machines which either directly encode or learn ethics and determine actions based on those ethics.

There are generally two approaches to implementing ethical behaviour in machines (Winfield and Jirotka 2018; Wallach 2008). A constraint-based approach (also known as top-down or rule-based), explicitly constraining the actions of an AI system under certain moral norms; and a training approach (also known as bottom-up or example-based), allowing the AI system to be trained to recognise and correctly respond to morally challenging situations. There might also be considered a mixed approach in which an AI system starts with a set of rules or values and modifies them into a system for discerning right from wrong. (Charisi et al. 2017).

Some work has been done on developing generalisable individual ethical decision frameworks combining rule-based and example-based approaches to resolving ethical dilemmas (Yu et al. 2018). Several ethical theories have been applied to Machine ethics: normative ethics (consequentialism, deontology and virtue ethics) (Yu et al. 2018; Carter et al. 2017), Rawls' veil of ignorance (Bowles 2018), or Habermas' discourse ethics (Mingers and Walsham 2010). Some models have been developed to consider data-driven examples (Balakrishnan et al. 2018); to reflect subjective preferences and ethical boundaries (Rossi and Mattei 2019; Loreggia et al. 2018); to represent ethical dilemmas (Anderson and Anderson 2014); Ethics Shaping, as a proposal to make reinforcement learners not only achieve the expected performance and the goals but also comply with ethical rules, using reward shaping and stochastic policy from human data (Wu and Lin 2018); or even a software "exoskeleton" that enhances and protects users by mediating their interactions with the digital world according to personalised data (Autili et al. 2019).

Research in approaches for implementing moral-decision making within AI systems is contributing to a more comprehensive interpretation of moral (Wallach and Allen 2012). For example, our human moral understanding is not semantically strict. Therefore, a certain degree of moral semantic flexibility is essential to morality itself, as we human beings understand and practice it. Hence, one question is how much semantic flexibility we should provide the AI systems with (Arvan 2018). Another big challenge is that of the consciousness and the necessary relationship between moral and first-person perspective (Nath and Sahu 2017). where there is no possibility of conscious experience, there is no subjective point of view, and, therefore, no way of taking such a system into account morally (Torrance 2013).[1] Some of the tools (such as the Moral Turing Test) for evaluating the moral performance of AI systems, would not be applicable

[1] For the purpose of this article we are not entering into the debate of moral responsibility vs moral accountability (see Floridi and Sanders 2004 or Bauer 2018). Our goal is not going into the attribution of those but rather to highlight when (human) moral control might be lost.

since that rests on "imitation" as a criterion for moral performance (Arnold and Scheutz 2016). Furthermore, a system which can pass an MTT may not give us a system which can provide satisfactory decisions in practical situations (Gerdes and Øhrstrøm 2015).

In the following section, we will describe a concept of practical and moral agency as a framework for drafting under which situations there might be a loss of moral control.

2 Practical and moral agency

2.1 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 and Alia 2015). At the same time, it can also be used to understand human decision-making—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 global environment.
- Desires[2]: states of the world–including herself–that are wished by the Agent.
- Intentions: future actions the Agent is committed to, to realise her desires given her beliefs.

From a logical point of view, 'beliefs' may be true or false, depending on the relationship between their content and the aspects of the world they intend to represent. 'Desires' have the opposite "direction of fit" (Tiberius 2015, p. 48): they do not intend to represent actual facts but to modify future facts, as to adjust them to what is desired. '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 particular desires, however, because of the commitment force involved, that transforms 'wish' into 'will', so to speak.

Intentions refer necessarily to the future. They are organised in 'plans' more or less precisely defined along time. The rationality of those plans requires them to be consistent with the Agent's beliefs and desires, and with her other intentions, in consequence with the beliefs and desires, those intentions intend to achieve.

2.2 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 realisation 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 herself morally, eventually modifying herself -the ancient theory of virtue as a habit built through exercise[3]).

The first meaning of moral appropriation (from person to act) constitutes the classical basis for 'moral responsibility': the Agent is also an author that can be called to respond of her choice by others. The idea is already found in Aristotle (2000, p. 1109b). Such moral responsibility may, in turn, become the ground for legal responsibility.[4]

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, p. 661) that generate some self-definition through choices. In consequence, it is not only a matter of behaving in specific ways but also of the internal makings that result in the Agent choosing that behaviour and the internal consequences of so doing. Those internal makings and consequences belong typically to the human psyche, and so 'morality' is a trait of humanity.

When applied strictly to people, Bratman's scheme allows us 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 iden-

[2] 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.

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

[4] 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).

tify all the beliefs in play in a moral decision[5], 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 entirely external states of the world, some others regarding relations with other people, and some related to her 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 seem 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 behaving as rational agents, or that we are behaving rationally but with beliefs or desires hidden even to ourselves.

Human morality understood as a concrete implementation of the practical agency has essential 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 'personalisation' of the AI-endowed machine, potentially useful for rhetorical purposes but prone to confusion.

In both cases, the discussion about assigning 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 to choose according to specific criteria in certain circumstances (see Wallach and Allen 2009).

2.3 Problems and messes

Moral rationality is not only a matter of optimising a particular 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 and 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 intends to maximise the achievement of a specific goal, given the available resources. In these situations, rationally choosing is a matter of optimisation.

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 a 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 the Agent, 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.

2.4 The basic question of the paper

As conceived up to the mid-twentieth century, machines merely constituted useful extensions of human action, increasing its range of possibilities in the most varied situations. They did not modify the morality of processes substantially. 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 'behaviour' 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 specific 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

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

the degree that is no longer acting entirely, the moral discussion loses its essential meaning.

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

3 Artificial 'reasoning'

3.1 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 et al. 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).[6]

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 the Universe to the sequent action in the Universe.

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 1976). Other approaches were provided, and one that carried 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 (Fig. 1).

A closer view of the Universe surrounding the rational agents would reveal a changing environment with heterogeneous sources of variation, as seen in Fig. 2:

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

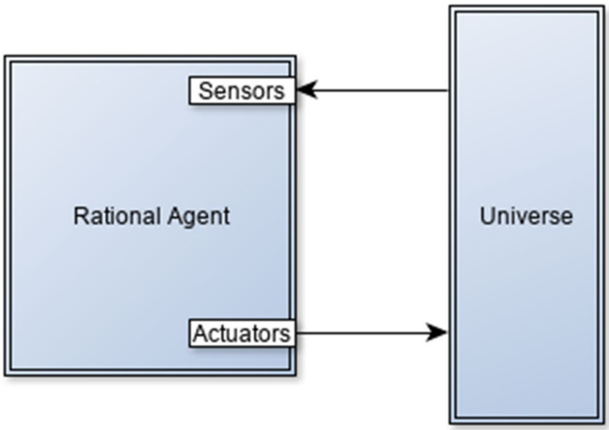

Fig. 1 Agent-Universe interaction

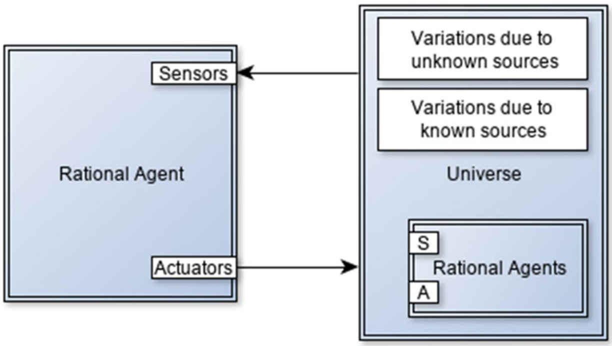
Sensors
Rational Agent
Actuators
Variations due to unknown sources
Variations due to known sources
Universe
S
Rational Agents
A

Fig. 2 Relationship between Agent and Universe

1. Variations due to transient conditions of known sources. "Known" here means that we know the origin of the variation, that we know its behaviour because it is produced by a phenomenon that follows some known rule (e.g. alternate current is known to toggle polarisation several times each second), or because we control that source of variation (e.g. a man switches on a light).
2. 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 also includes any inaccuracy that the Agent may experiment while acquiring the external signal with the sensors.
3. 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.
4. 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

[6] 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.

because each one is guided by a "utility function", that expresses the value that Agent aims to maximise 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.

3.2 Expert knowledge

The relevant question to be answered in this paper regards the behaviour of the agents. The simplest kind of behaviour 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 simply and transparently, the knowledge of the experts in the subject matter (thus the name). The resulting Expert System aimed to make a "twin" of how 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 could not cope with so many circumstances as the sources of variation could produce. Even if it could work with all the past and present conditions, it might fail in the future. 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.

3.3 Learning agents

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

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

To assess which is the best action to adopt, the Rational Agent is endowed with a "utility function". This function is essentially an internalisation of a performance measure. If the internal utility function and the external performance measure are in agreement, then an agent that chooses actions to maximise its utility will be rational according to the external performance measure. Figure 3 represents this separated internal function of the AIS.

Utility functions are useful to manage some conflicting situations (Russell et al. 2010, p. 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 trade-off. Second, when there are several goals that the Agent can aim for, none of which can be achieved with certainty, the utility function 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 maximises 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.

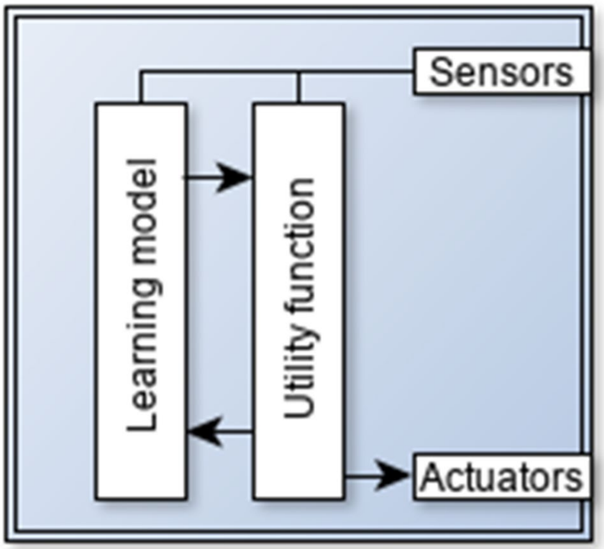

Fig. 3 Agent with Model and Utility function

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 the account, collaboratively or adversely. This makes the assessment of the utility function much more complicated.

3.4 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 of this capacity to understand it better. It is called "Machine Learning", as a generic name.

In Machine Learning, models are usually grouped 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).

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

- Clustering: used to group similar samples. A common useful measure for the goodness of the clustering is called "entropy". The clustering techniques aim to minimise the internal entropy in clusters while maximising 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 on 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: (Sutton et al. (2018) 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 maximise the reward. Incipient as it is, Reinforcement Learning brings the possibility to build AIS that learn their utility function directly from the Universe where they operate.

3.5 Deep learning

Any of the systems previously described become overwhelmed in many real situations, even those that are very simple for the human brain, like identifying an image or understanding a speech, among others. The jump into more complex achievements come from the ability of these systems for dividing the target into many layers, evaluating different characteristics of the input processing every one of them in parallel, and feeding the results of a process to the next layer. This chain recovers the original compositional hierarchies and allows a higher abstraction, e.g., analysing a set of pixels with different colours, bright, etc., and as a result labelling the set as a "dog", or a set of words and through several processes of recognition of entities, lemmatisation, etc.

This process mimics the way of working of the human nervous system, in which different parts of the brain specialise in the characteristics of the input, but process them as a whole. They bring the ability to reduce the complexity of the inputs (e.g. pixels) to more abstract entities (e.g. dog), that allows working with them in logical and moral decisions. Deep learning can be used in supervised and unsupervised learning tasks, as stated above. The exponential advantages of deep, distributed representations help to tackle the exponential challenges of the curse of dimensionality when complex problems are the target. Deep learning applications range from Artificial Vision (Convolutional Neural Networks), Natural Language Processing (Recurrent Neural Networks) and GAN (Generative Adversarial Networks) to generate images and sounds that look "real".

Deep learning can also be applied in reinforcement learning problems, and the technique is usually known as Deep Reinforcement Learning (Deep RL). Deep RL is a particular type of RL, with deep neural networks used for state representation or approximation for the value function. This is one of the most recent techniques, the first example in the literature appeared in (Mnih et al. 2015). The system can learn an action strategy using inputs from more complex sources of data.

3.6 Two problems for the moral agency in the 'reasoning' of AIS

The moral agency requires some intelligence of the situation, 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, 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 modelled using Bell curves, adequate for small distances from the mean but grossly inadequate for extreme events when very few cases for training AIS were available.

Going out of limits can also be understood qualitatively: AIS consider only the situation for which they have sensors. In the terms we presented in 2.3, they can 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 useful, and in consequence which output would be adequate. The AIS may be '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 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 the 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 the need for external assistance.

Taking moral decisions in systems that include AI 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. The area of "Explainable AI" has attracted much interest from both academic and practitioners (Barredo et al. 2019)

4 Data quality

4.1 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 the 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. Thus, data themselves play an essential role in the definition of the capabilities of any such algorithm. As a matter fact the significant 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, p. 62).

Data sets are indeed a crucial part of AI systems, and as such, they are receiving increased attention from researchers and companies (Vakkuri 2018). Both supervised and unsupervised learning depends on training data, i.e., known datasets for practising 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 the 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 et al. 2018). Furthermore, to be usable, data must be treated and conditioned with tools. Therefore, the "rawness" of data may be disrupted in various manners (Ekbia 2015), which might introduce further distortion in the data sets.

For 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 Fig. 4):

4.2 Origin of data

Every company, in a declared or undeclared manner, benefits from capturing more data. Machine learning is only useful 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 marginalised by this need of scale (McQuillan 2018), left apart from the moral consideration about respect for human dignity, which requires that all people are treated with the respect due to them as individuals, rather than merely as data subjects.

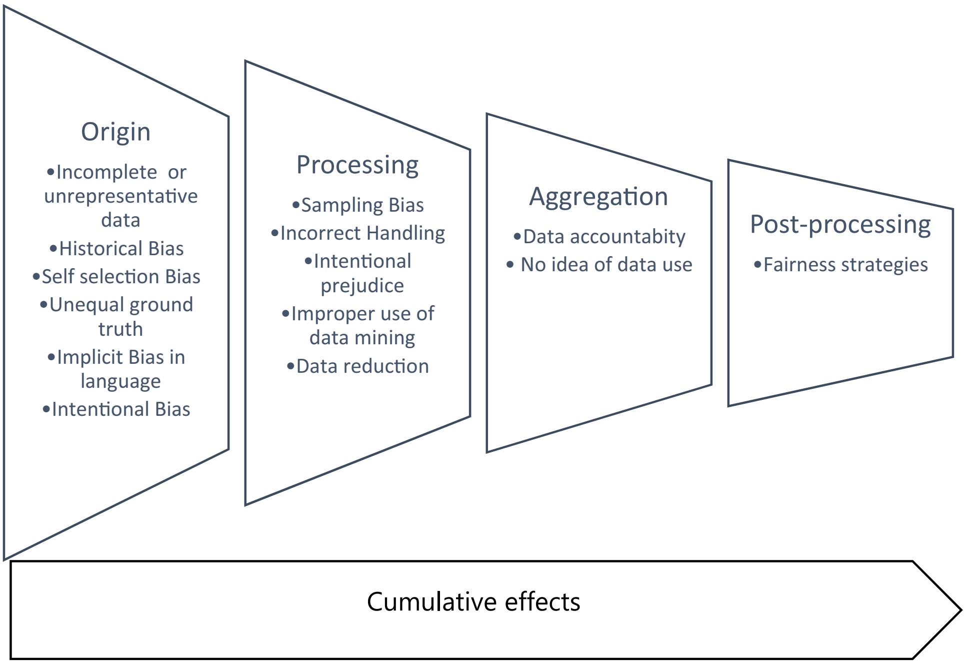

Fig. 4 Classification of potential issues related to data sets

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 adequately selected: "it is an error to assume 'people' and 'twitter users' are synonymous" (Boyd and Crawford 2012, p. 668).

Data might be tainted with a 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 several kinds of biases, 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 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 the language (Caliskan et al. 2017), which is to be considered when utilising word embeddings or semantics representation of words in AI systems.

There might also be 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).

4.3 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 implicit assumptions (Bowles 2018). Pre-processing of data might include data integration, data cleaning, data transformation, data reduction and discretisation (Mnich 2018). Several issues might arise at this pre-processing stage: sampling bias, causing that part of the population to be 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 to exclude certain groups of persons, by explicit manipulation of the data.

There is another issue related to the pre-processing of data: improper use of data mining techniques, that is analysing 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, p. 80).

4.4 Aggregation of data

'Big data' are not merely 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.

Concerning this, there is a risk of non-fully considering the potential data use, based on the false 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 especially challenging when using data output as data input in the "mess" architecture described above.

4.5 Post-processing of data

To countereffect some of the issues mentioned above, efforts have been made 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).

4.6 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 particular 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 handlings, such as improper labelling. In those cases, the potential biases propagate with unforeseeable consequences and loss of control over the moral decision.

To fight data bias, AI should always be applied transparently, to understand, monitor and suggest improvements to algorithms; it is also suggested to include diversity among AI developers, 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 a 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 to minimise 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, or Fairness-flow by Facebook, to name a few. (AI Now 2018).

Nevertheless, there are several 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, to apply the proper fairness strategy to each situation? Furthermore, the proliferation of observational fairness methods through algorithmic treatment, which are not entirely stable, as explained above, might provide a sense of false security (AI Now 2018).

5 Decision making

5.1 Information and action

AIS are trained and fed with data (Sect. 4) and operate on them 'intelligently' (Sect. 3) 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 a third agent(s) or an action executed by itself.
- Action consisting of physical actions–for example in a robot—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. However, the provision of information for its decision making may be quite complex in design, including other operational units networked in different ways.

5.2 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 faster and execute decisions (being the case) almost instantaneously. They can act on patterns that people would not 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 always produce the same output.
- AIS are less flexible than people. They can only calculate on the input they are programmed to consider. At the same time, people can 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 (Sect. 2).

AIS are being used in three different ways concerning the decision making that leads to 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): always controlled by an AIS, that makes all decisions in all circumstances.

All these systems have common problems related to information, recounted in Sect. 4. 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 (Sect. 3), the decision will also be, in the sense that it will be impossible to 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, for 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 sure '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 particular 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.

5.3 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.

This is the least problematic use of AI machines for decision making. The human decision-maker needs not to 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, utilise her own heuristics…

The problems related to the 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 the usual human heuristics. Those solutions can be provided by a behavioural 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.

[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.

- 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 importantly, the institutional authority that adopted that precise DSS.

5.4 Semi-autonomous systems (SAS)

The SAS operate autonomously in many situations but can 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 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 of 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. In addition, 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 decisive 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 the 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 the machine and human decision-makers (see "5.2" 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. In contrast, the human decision-maker may add her flexibility, professional experience and general ability to place the situation in a broader 'mess' context.

5.5 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 making all the moral decisions beforehand, the AS may operate in an eventually unpredictable—undecipherable—way.

The problem of traceability (transparency) described in 3.6, 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 regain control over the functions the AIS was controlling.

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 is one reason, among others, why SAS are generally considered preferable to AS.

5.6 Information integration

Salgues (2018, p. 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 realise 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.

6 Conclusions

In this article, we have offered an elemental mapping of the technological architectures that support AIS, under the specific focus of the 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 summarise our findings, we present them according to the essential elements of human morality that may be affected (see "2.1"):

6.1 Beliefs

AIS collect and process the information relevant for maximising 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 "3.4").

In all three cases, it depends on the Universe of cases fed to the AIS. 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 4.2);
- the processing of data: sampling bias, incorrect labelling, intentional manipulation, improper use of data mining techniques (see 4.3);
- the aggregation of data from different origins in time, place, the procedure of recollection… each one blindly involving its own options (see 4.4).
- the post-processing of data with bias-cleaning algorithms, which introduce another layer of moral criteria often unknown to the user (see 4.5 and 4.6).

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

6.2 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 "2.3"). 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 intentional management of the 'mess' where the problem makes sense, which is a typical work of a human moral agency.

Having an AIS 'solution' available may foster 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 intend 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 "3.1") poses a major issue relating to 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 maximise. 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.

6.3 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 "5.5").

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 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 "5.3").

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 "5.4").

6.4 Perspectives

Moral agency requires a special implementation of the BDI logic that corresponds to the human psyche and has its characteristic openness to the world and to itself (2.3). If an AI-endowed practical agent somehow nullifies some of its elements, 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 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 summarised in this paper but also about the question of moral agency through complex networks that include AIS and human operators.

References

Anderson M, Anderson SL (2014) GenEth: a general ethical dilemma analyser. In Twenty-Eighth AAAI Conference on Artificial Intelligence.

Aristotle, Crisp R (2000) Nicomachean ethics. Cambridge University Press, Cambridge

Arnold T, Scheutz M (2016) Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. Ethics Inf Technol 18(2):103–115

Arvan M (2018). Mental time-travel, semantic flexibility, and AI ethics. AI & SOCIETY, pp. 1–20

Autili M, Di Ruscio D, Inverardi P, Pelliccione P, Tivoli M (2019) A software exoskeleton to protect and support citizen's ethics and privacy in the digital world. IEEE Access 7:62011–62021

Balakrishnan A, Bouneffouf D, Mattei N, & Rossi F (2018) Using contextual bandits with behavioral constraints for constrained online movie recommendation. In IJCAI (pp. 5802–5804)

Barredo AA et al. (2019) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. arXiv preprint arXiv:1910.10045

Bauer WA (2018) Virtuous vs utilitarian artificial moral agents. AI & SOCIETY, pp. 1–9

Boddington P (2017) Towards a code of ethics for artificial intelligence. Springer, Cham

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

Boyd D, Crawford K (2012) Critical questions for Big Data. Inform Commun Soc 15(5):662–679. https://doi.org/10.1080/1369118X.2012.678878

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

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

Caliskan Aylin, Bryson Joanna J, Narayanan Arvind (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186

Carter SM, Mayes C, Eagle L, Dahl S (2017) A code of ethics for social marketing? Bridging procedural ethics and ethics-in-practice. J Nonprofit Public Sect Mark 29(1):20–38

Charisi V, Dennis L, Fisher M, Lieck R, Matthias A, Slavkovik M, Sombetzki J, Winfield AF, Yampolskiy R (2017). Towards moral autonomous systems. arXiv preprint arXiv:1703.04741

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

Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, Bowman T, Suri VR, Tsou A, Weingart S, Sugimoto CR (2015) Big data, bigger dilemmas: a critical review. J Assn Inf Sci Tech 66:1523–1545. https://doi.org/10.1002/asi.23294

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

Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14(3):349–379

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, 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, Oxford

Gerdes A, Øhrstrøm P (2015) Issues in robot ethics seen through the lens of a moral Turing test. J Inform Commun Ethics Soc

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

Greene D, Hoffmann AL, Stark L (2019) Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences

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 PT, Adams KM (2017) Systemic decision-making fundamentals for addressing problems and messes. Springer, New York

AI HLEG, High-Level Expert Group on Artificial Intelligence (2018) Draft ethics guidelines for trustworthy AI. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=57112

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399

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

Loreggia A, Mattei N, Rossi F, Venable KB (2018) Preferences and ethical principles in decision making. In 2018 AAAI Spring Symposium Series

Markham AN, Tiidenberg K, Herman A (2018) Ethics as methods: doing ethics in the Era of big data research—introduction. Soc Media Soc. https://doi.org/10.1177/2056305118784502

McCarthy (1958). "Programs with common sense." Proceedings of Teddington Conference on the mechanisation of thought processes

McQuillan D (2018) People's councils for ethical machine learning. Soc Media Soc 4(2):2056305118768303

Metcalf J, Emily FK, Danah B (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, Broersen Jan, Herzig Andreas (2015) BDI logics. In: Ditmarsch HV (ed) Handbook of epistemic logic. College Publications, London

Mingers J, Walsham G (2010) Toward ethical information systems: the contribution of discourse ethics. Mis Quart 34(4):833–854

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

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533

Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21

Nath R, Sahu V (2017) The problem of machine ethics in artificial intelligence. AI Soc 35(1):103–111

Plato et al (2018) The republic. Cambridge University Press, Cambridge

AI NOW Report (2018) Artificial Intelligence Institute. New York

Rossi F, Mattei N (2019). Building ethically bounded AI. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 9785–9789)

Russell SJ, Norvig P, Davis E (2010) Artificial intelligence: a modern approach, 3rd edn. Prentice Hall, Upper Saddle River

Salgues B (2018) Society 5.0. Wiley, Hoboken

Shortliffe EH (1976) Computer-based medical consultations: MYCIN. Elsevier, Amsterdam

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

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

Sutton RS, Barto AG (2018) Reinforcement learning: an introduction, 2nd edn. MIT Press, Cambridge

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

Tomasello M (2018) Precís of a natural history of human morality. Philos Psychol 31(5):661–668

Torrance S (2013) Artificial agents and the expanding ethical circle. AI & Soc 28(4):399–414

Vakkuri V, Abrahamsson P (2018) The key concepts of ethics of artificial intelligence. In 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1–6). IEEE

Wallach W (2008) Implementing moral decision making faculties in computers and robots. AI & Soc 22(4):463–475

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

Wallach W, Allen C (2012) Hard problems: framing the chinese room in which a robot takes a moral turing test. University of Birmingham, AISB/IACAP, p 5

Winfield AF, Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos Trans R Soc A 376(2133):20180085

Winfield AF, Michael K, Pitt J, Evers V (2019) Machine ethics: the design and governance of ethical AI and autonomous systems. Proc IEEE 107(3):509–517

Wu YH, Lin SD (2018) A low-cost ethics shaping approach for designing reinforcement learning agents. In Thirty-Second AAAI Conference on Artificial Intelligence

Yu H, Shen Z, Miao C, Leung C, Lesser VR, Yang Q (2018) Building ethics into artificial intelligence. arXiv preprint arXiv:1812.02953

Zilberstein S (2015) Building strong semi-autonomous systems. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.