AI: from rational agents to socially responsible agents

Antonio Vetrò[1,2], Antonio Santangelo[1], Elena Beretta[1], Juan Carlos De Martin[1]

[1] Nexa Center for Internet & Society, DAUIN, Politecnico di Torino, Italy

[2] Future Urban Legacy Lab, Politecnico di Torino, Italy

{name.surname}@polito.it

Structured Abstract

Paper category: Conceptual paper.

Purpose (mandatory) The paper analyses the limitations of the mainstream definition of Artificial Intelligence (AI) as a rational agent, which currently drives the development of most AI systems. The authors advocate the need of a wider range of driving ethical principles for designing more socially responsible AI agents.

Design/methodology/approach (mandatory) The authors follow an experience-based line of reasoning by argument to identify the limitations of the mainstream definition of AI, which is based on the concept of rational agents that select, among their designed actions, those which produce the maximum expected utility in the environment in which they operate. Then, taking as an example the problem of biases in the data used by AI, a small proof of concept with real datasets is provided.

Findings (mandatory) The authors observe that biases measurements on the datasets are sufficient to demonstrate potential risks of discriminations when using those data in AI rational agents. Starting from this example, the authors discuss other open issues connected to AI rational agents and provide a few general ethical principles derived from the experience of the White Paper *Artificial Intelligence at the service of the citizen* (Agid 2018).

Originality/value (mandatory) The paper contributes to the scientific debate on the governance and the ethics of Artificial Intelligence with a novel perspective, which is taken from an analysis of the mainstream definition of AI.

Keywords: Artificial Intelligence, Data Ethics, Digital technologies and society

1. What kind of rationality for AI systems?

The expression "Artificial Intelligence" is gaining considerable attention from both private and public sector (Runkin, 2018) (Roy, 2018). The hype is very high and, as it often happens in such situations, all this attention has generated confusion, even among experts, who refer to Artificial Intelligence to talk about very different things.

We refer to AI following the mainstream definition of Russell and Norvig (Russell and Norvig, 2010): it is "*the study of designing and building intelligent agents* (p.30), where "agent" is "*anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators*" (p.34). An intelligent agent "*takes the best possible action in a situation*" (p.30), i.e. it is a rational agent the one which, for each possible percept sequence, is supposed to "*select an action that is expected to maximise its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has*" (p.37).

An advantage of this definition is that the few concepts above are the building blocks for designing AI systems with scalable complexity, e.g. from a "simple" vocal translator to an autonomous vehicle. However, this definition of AI is based on a very precise, and in a sense narrow, vision of the concept of intelligence, which is bound to a particular type of rationality. In fact, if the actions undertaken by an agent must always maximize a performance measure, it is clear that the functionality and the effectiveness of such actions are strictly dependent on the form of knowledge the agent itself incorporates: an action is always the consequence of a certain vision of the world, of the world's rules that are considered to be true and for this reason are embedded in the form of algorithms elaborating data, and of a precise conviction about what the world should become according to that logic.

There is a vast amount of evidence showing that designing and building AI agents according to such deterministic perspective is producing relevant negative social effects.

Recently, the investigative website ProPublica discovered the COMPAS algorithmic tool - Correctional Offender Management Profiling for Alternative Sanctions - widely spread in the American criminal justice to prevent recidivist behaviors, was biased against black defendants, revealing the tool was assigning them a higher risk rate generally (Angwin at al. 2016); Latanya Sweeney (2013) has highlighted that delivery of ads by Google AdSense was biased in the sense that in a Google search for an individual's name an arrest record was suggesting by algorithm on the basis of racial association, and Cathy O'Neil underlined a large-amount of case-studies in which people are subjected to racial, gender, or any other kind of discrimination, in AI ground (O'Neil 2016).

In addition, there are already many examples, in this regard, from AI agents that help financial institutions decide to which category of people to lend money, and that are based on the idea -implicitly embedded in the code and in the data used by the software- that it is better to favor white, educated citizens residing in certain specific areas of the cities, and especially males (Credit Suisse, 2017); or other examples that include AI agents deciding (or recommending) to grant probation to prisoners, which once again favor individuals belonging to certain ethnic groups, or targeting to men more than to women job offers that are more economically advantageous (Spice, 2017). Evidently, the "forms of knowledge" on which the algorithms that "animate" these machines are based, are the result of databases (or, in the simplest of cases, statistical surveys), which, even if accurate, they represent certain distortions of our society. So the question is: do we want these distortions to increase and to be perpetuated by our Artificial Intelligence tools, or do we prefer to create instruments that may help us diminish the unjust situations we live in?

In this sense, the scientific world is taking important steps to include other perspectives, such as the ethical one[1], at the center of Artificial Intelligence programming, in order to avoid giving rise to a world in which we can design certainly effective and high-performative AI agents, but at the same time let them decidedly unfair, and in our place. This paper is part of this community effort, and we advocate the need of a wider range of designing principles for AI agents, which goes beyond the perspective of the mainstream definition of AI. The remainder of the paper is structured as follows: we focus on the problem of bias in the data in Section 2, by identifying some measures. In Section 3 we apply the bias measures on three real datasets. This proof of concept leads us to a conceptual reflection on the need of socially responsible agents instead of rational agents (Section 4), and provide a few high-level policy guidelines, extracted from the White Book Artificial Intelligence at the service of the citizen (Agid 2018). We summarize and draw conclusions in Section 5.

2. Bias of the forms of knowledge governing AI

The problem of bias in the data used by AI systems is well represented in the following excerpt of Cathy O'Neal's book "Weapons of Math Destruction" (O'Neal, 2016):

"if the admission models to American universities had been trained on the basis of data from the 1960s, we would probably now have very few women enrolled, because the models would have been trained to recognize successful white males."

The observation made by O'Neal entails an important, more general, reasoning: not only how AI collects and elaborates data has ethical consequences, but, before that stage, also the input data properties (*percepts*, in the terminology analyzed in the previous section) are connected to important ethical interdisciplinary issues. The characteristics of the "forms of knowledge" involve ethical issues (Floridi and Taddeo, 2016), and those problems propagate downwards throughout all subsequent phases of the data life cycle in AI systems, until affecting the output, i.e. the decisions or recommendations made by the software. Therefore, certain data characteristics may lead to discriminatory decisions and therefore it is important to identify them and show the potential risks. We take as reference two characteristics of input data: disproportions and collinearity.

2.1 Disproportionate datasets

AI systems work on the base of large amount of historical data, very often elaborated with machine learning models. Problems of fairness and discrimination may arise due to disproportionate datasets, which lead to disproportionate results, generating problems of representativity when the data are sampled -thus leading to an underestimation or an overestimation of the groups - and of imbalance when the dataset used has not been generated using the classical sampling methods. Simple random sampling - which is the most widely used method in statistical surveys - requires that the probability of sample extraction is known and not zero, and that not only each element but also each combination of elements (of equal number) has the same probability of being extracted. A biased sample leads to biased estimates. For this reason, statistical sampling is a fundamental step. However, in the era of Big Data, many of the data used today have not been generated using probabilistic sampling, but are rather selected through non probabilistic methods (very often acquired from third parties, or with opportunistic methods, thanks to the pervasiveness of digital technologies), which do not provide to each unit of the population the same opportunity to be part of the sample; this means that some groups or individuals are more likely to be chosen, others less. Representativity is a property of the

[1] See for example of the joint initiative of Berkman Klein Center of Harvard University and MIT Media Lab, who are leading a $27 million program to study ethics and governance of Artificial Intelligence. (https://cyber.harvard.edu/research/ai: last consultation 29 August 2018)

outcome of the extraction process, which itself has randomness as its property. For this reason, it is essential to keep this aspect under control in non-probabilistic samples.

In general, solutions relating to demographic or statistical parity are useful in cases where there is no deliberate and legitimate intention to differentiate a group considered protected, which would otherwise be penalized (Dwork et al., 2012). It should therefore be borne in mind that the solutions vary according to both the nature and use of the data. Take as an example a type of analysis that includes in its attributes the individual income. If the choice to include in the sample only individuals with a high income is voluntary, no representativity problems arise, since the choice of a given group is based on the purposes of the analysis. However, if the probability of being included in the sample is lower as the income is lower, the sample income will on average be higher in the overall income of the population.

2.2 Collinearity

In statistics two variables X_1 and X_2 are called collinear variables when one is the linear transformation of the other and therefore there is a high correlation. Collinearity is a group phenomenon involving at least two regressors and which may affect, in different extents, different groups of regressors. In general, there are always relationships between regressors that involve a certain degree of linear dependence, but it is good practice to consider the correlation value 0.9 as the limit beyond which singularity or almost singularity in the matrix of regressors is observed; over this threshold the estimation of parameters in Ordinary Least Squares are to be considered not reliable.

In general, the main causes of collinearity are due to data collection techniques, such as similar measurement errors on different regressors; spurious correlations; inconsistency of a regressor data with the model specification, e.g. when using a higher than necessary polynomial; or application of a model to a small number of cases. The attempt to contain the negative effects is mostly due to the fact that collinearity damages the estimates of parameters and their precision. To prevent this effect some researchers adopt a naïve approach that precludes the use of sensitive attributes such as gender, race, religion and family information, but may not be effective in case of multicollinearity. The use of geographic attributes, for example, is reported to be unsuitable when the use of protected data is to be foreclosed, because it easily leads to tracing protected attributes, such as race (Lepri et al., 2017). Hardt (Hardt et al. 2016) points out that the condition of non-collinearity requires that the predictor (Y') and the protected attribute (A) are independent conditional on Y - e.g., the variable to be predicted, income, must be independent of the gender variable. In practice, it is encouraged to use features that allow to directly predict Y, but prohibits abusing A as a proxy for Y.

Another common error is "mistake correlation with causation"; a high entropy dataset can induce thinking that the large number of features is sufficient to explain causality. Cause-effect ratios are often confused with correlations when features are used as proxies to explain variables to be predicted. For example, the IQ test is a test that measures logical-cognitive abilities, but if used as a proxy to select the smarter students for admission to a university course, it would almost certainly reveal itself as an imperfect proxy, since intelligence is a too broad concept to be measured by a number only. As a consequence, the test of the IQ is not sufficient to explain the variable to be predicted.

To avoid the risks mentioned above, the following thresholds, defined on the base of literature and experience, are useful to identify cases of collinearity:

1. correlation values higher than 0.9 should be avoided;
2. the absence of high correlations does not exclude collinearity; it is therefore always good to also consider the value of R^2, in the case of $R^2 = 1$, we are in presence of multicollinearity;
3. in case of collinearity there is no increase in the explained deviance which is certainly attributable to the effect of a specific regressor.

In addition, an effective method to identify collinearity is to calculate the Variance Inflation Factors that indicate how much parameter variability depends on the regressors. VIFs are calculated in this way:

$$VIF_i = \frac{1}{1 - R_i^2}$$

If the i-th regressor is not linearly linked to the others, $R_i^2 = 0$ and VIF will be equal to one. High levels of VIFs indicate the presence of a relation of linearity: it is commonly assumed that for $VIF(\beta_i) > 10$ multicollinearity is strong.

Finally, since correlation measurements can only be used for quantitative variables, the degree of dependency between categorical data is measured using the estimation of Pearson residuals, which is a commonly accepted measure of discrepancy between observed and expected values (Zeileis,2007).

3. Measures on datasets

3.1 Overview of data sets used

We applied the metrics defined above to the following three datasets, each referring to a different application domain. Table 1 shows to which datasets which measure was applied.

Credit card default dataset (Lichman, 2013)**.** This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The dataset is composed by 24 variables, of which four demographic ones that can be considered as protected attributes (sex, age, education, marital status).

COMPAS Recidivism racial bias dataset (Larson et al., 2016)**.** Data contains variables used by the COMPAS algorithm in scoring defendants, along with their outcomes within two years of the decision, for over 10,000 criminal defendants in Broward County, Florida. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an algorithm used by judges to assess the probability of recidivism of defendants. Three subsets of the data are provided, including a subset of only violent recidivism, as opposed to, e.g. being re-incarcerated for non-violent offenses such as vagrancy or Marijuana. The original dataset contains 28 variables, of which eight are considered protected attributes: Last name, first name, middle name, sex, race, date of birth, spoken language, marital status. The dataset is well-known because of a study of the U.S. non-profit organization Pro Publica (Angwin et al., 2016) that showed that the COMPAS algorithm is distorted in favor of white individuals, thus exposing black people to a risk of distorted recidivism, because it would be higher than it actually was.

Student alcohol consumption dataset (Cortez and Silva, 2008)**.** The data were obtained in a survey done by students of mathematics and Portuguese language courses in secondary school. It contains social, gender and study information about students. Two datasets are provided: The one containing the students of the mathematics course contains 395 observations, the one relating to the Portuguese language course contains 649 observations. Both contain 33 variables, most of which are protected attributes describing demographics, such as school, context of belonging (urban, rural), family indications, etc.

	Disproportion	Collinearity
Credit card default dataset	✓	✓
COMPAS Recidivism racial bias dataset	✓	
Student alcohol consumption dataset		✓

Table 1 Measures and datasets

3.2 Measures on datasets

3.2.1 Credit card default dataset

The field of creditworthiness often appears in the literature alongside issues related to ethical decisions (Yang 2006, Rice et al. 2017). Recently, some studies have shown that access to credit for black people is modulated by certain attributes such as race, rather than by information about the payer's status (NY Times, 2018) (Bartlett et al. 2017, Chen 2018). The dataset that we use does not contain the protected attribute race, however it contains other personal information that can be used in a discriminatory way if applied to assess creditworthiness, such as gender and level of education.

Disproportion Figure **1** reports the frequency of variables gender, marital status, age, education, expressed as a percentage for each of their categories. The data shows that 60% of individuals are women, 46.7% of individuals have attended university, the age group most represented is that of 25 to 40 years, the proportion of married individuals is the same for single individuals.
Although we do not have information neither on the real frequencies of protected attributes in the source population nor on the sampling method used (if any), the results of the analysis of disproportions suggests to use the age variable with caution: in fact the variable age shows a more considerable disproportion compared with the other protected attributes, exposing a potential risk of discrimination (e.g., if the dataset is used to automate decisions or recommendations on the capability to repay a debt, and attribute age is one of the predictors).

Collinearity We perform the analysis for each protected attribute in the Credit Card default dataset, in relation to default payment (1 = yes, 0 = no). We report on Figure **2** the mosaic plots[2] for the attributes education, marital status and gender: blue indicates cases in which there are more observations in that cell than would be expected under the null model of independence between attribute education and attribute default payment; red means there are fewer observations than would have been expected; eventually, grey indicates that observations are coherent with the assumption of independence. Figure 2 shows that:

- default payment is highly correlated to the education level, for all its levels;
- the correlation between the protected attributes and the default payment variable is significant for the gender variable (both male and female);
- the correlation is significant for the marital status variable in correspondence with the default payment group = yes.
- In addition, Pearson residuals[3] show that the most correlated categories are: the education variable and the male, both in correspondence with default payment = yes.

As a consequence of the analysis, the identified correlations should be taken into account when using the dataset in an algorithm that supports or automate decisions.

3.2.2. COMPAS Recidivism racial bias dataset

Disproportion As reported above, previous research has shown that the data in the COMPAS dataset is unbalanced in favor of white people. Table **3** shows the variability in race attribute, which is the underlying reason of the findings of the previous study: the highest levels of reoffending are observed in black individuals. Moreover, 33.22% of the dataset's observations

[2] A mosaic plot is an area proportional visualization of a (possibly higher-dimensional) table of expected frequencies.

[3] Pearson residuals are widespread in statistic domain to study the linear relationship among two variables. See more at: https: // en. wikipedia. org/ wiki/ Pearson_correlation_ coefficient and at:https: // en. wikipedia. org/ wiki/ Pearson\%27s_ chi-squared_ test

refer to white people, while 53.45% refer to black people, indicating that there may be an overestimation of the race attribute - against black people - which would contribute to the estimation of recidivism (confirmed by follow-up analyses showing a highly dependence between that the race variable and the level of recidivism).

3.2.3 Student alcohol consumption dataset

The dataset is composed of 33 variables, principally qualitative; 649 observations for the dataset referring to students of Portuguese, 395 for that referring to students of mathematics.

Collinearity We randomly chose 4 quantitative variables to predict workday alcohol consumption and calculated the Variance Inflation Factor for each of the regressors, and report results on Table **2**. The variables indicate: i) number of school absences (numeric: from 0 to 93); ii) current health status (numeric: from 1 - very bad to 5 - very good); iii) quality of family relationships (numeric: from 1 - very bad to 5 - excellent); iv) age. The average of VIF is equal to 1.02, therefore among the variables considered a relationship of collinearity is only moderate. However, we observe that while some attributes in some contexts are considered protected, in others are essential to avoid situations of risk or damage; in the case of alcohol abuse among students, personal information are useful to identify areas of intervention and define appropriate social policies. We underline once again how the intended scope of the AI plays a fundamental role in the choice of considering some attributes as protected or not.

Ethnic group	High	Low	Medium	N/A
African-American	3400	3369	3010	12
Asian	9	50	12	0
Caucasian	943	3554	1579	10
Hispanic	191	945	315	0
Native American	15	26	16	0
Other	56	653	150	1

Table 3. Distribution of ethnic groups within the COMPAS database, by level of risk of recidivism

	Absences	Health	Fam Rel	Age
VIF	1.02	1.01	1.01	1.03

Table 2. Variance Inflation Factor for selected attributes in student alcohol consumption dataset

4. For Artificial Intelligence as a socially responsible agent

The exemplary measurements reported in the previous section highlight that when the biases of the world in which we live are inserted into the AI agents designed only to maximize a performance measure, certain injustices can only be perpetuated and exacerbated. In the recent years, an animated debate on this subject has risen in the scientific community and in the civil society, as the use of the so-called "Big Data" has shown all its potential in different areas, including that of Artificial Intelligence. As we have seen before, in this field a relevant problem is that of unbalanced data sets, which overestimate or underestimate the weight of certain variables - often, once again, related to gender or linked to the belonging of individuals to some minorities - in the reconstruction of the cause-effect relationship necessary to predict events, as was the case with some algorithms used by the American police to prevent crimes (Lum and Isaac, 2016). The authors demonstrated that the algorithm used by police patrols for predicting future drug crimes, were fed with data that were under-representative of the white consumers of drug. As a result, predictions of the software constantly pointed to areas of the cities where non-white people resided, and police would follow those recommendations to focus crime prevention activities. Arrests would then be concentrated in those areas, and on the non-white people, creating a feedback loop that reinforces the initial bias.

In addition, there are cases where data biases (or those embedded by algorithms) are not just a simple reflection of the world we live in, but can be injected during the agent training process (Barocas and Selbts, 2016), as it happens with the techniques of supervised learning - at the moment among the most widespread - where machines must be instructed to carry out their calculations and need data that are "noted" by human beings. However, also data annotated by humans are not free from bias: it has been reported, for example, that differences in gender or ethnic and social origin can produce different biases in the evaluation of the meaning of an image or of a concept (Bencke, 2016) (Crawford, 2016).

The ethical issues raised by the functioning of AI go well beyond the composition of its databases.

Then, Artificial Intelligence poses problems of transparency and openness, since it is often not possible to determine what are the data on which it bases its operation, nor the architecture of its algorithms, which are covered by industrial secrecy. This can be dangerous in many areas. For example, in the world of employment, perplexity is beginning to arise over the use of Artificial Intelligence tools in the selection and management of personnel, the mechanisms of which are unknown to employees and intermediate bodies. But consider also the dystopian scenarios of the adoption of "opaque" machines by the State, which would administer its power without allowing citizens to control its actions. For this reason, in countries like France, an attempt is made to pursue a policy linked to the promotion of open data and of the open code (Di Cosmo and Zacchiroli, 2017). However, there are still situations in which transparency and openness do not imply two other desired properties of AI: explainability and understandability. It is the case for example of neural networks, whose algorithms of calculation could not be completely reconstructed even by their programmers, generating what is called in the jargon "black box effect" (Ritter et al., 2017) (Knight, 2017).

This issue has been regulated by the new General Data Protection Regulation at Article 22, that provides a general prohibition of solely automated decision-making (that means with no significant human intervention in any phase of the data processing) with legal effects on the individual; furthermore, Article 29 of the "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" specifies that transparency criterion is mandatory also in cases of high complexity of the technology concerned, as the transparency conditions require that detailed explanations or disclosures of the whole algorithm are not to be disclosed, but rather the underlying logic in order to clarify the criteria leading to a particular decision.

Finally, the counterbalance of openness and transparency is the need to protect the privacy of individuals, leading to the setting of boundaries beyond which transparency cannot be pursued. One of the typical nodes, in the field of Artificial Intelligence and not only, is for example that of the so-called "mosaic effect", linked to the secondary use of certain data, very frequent in health research, which is not easy to be predicted from the beginning and which, for this reason, makes informed consent complex to implement. A consensus that must be called into question also to establish in which situations each of us can refuse to be subjected to "treatment" by means of Artificial Intelligence. This issue has been regulated by the new General Data Protection Regulation at Article 22, that provides a general prohibition of solely automated decision-making (that means with no significant human intervention in any phase of the data processing) with legal effects on the individual (see also Art. 29 of the "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679").

In the light of the exemplary issues summarized above, we can conclude that although a deterministic approach to the design and building of AI agents can give the impression of giving "scientific and objective" bases of their decision, such logic does not avoid the risk of discriminating decisions. A more comprehensive approach is needed, which should take into account, alongside the definition of Artificial Intelligence as a "rational agent", another one in parallel: that of a "socially responsible agent". AI should be rational in augmenting social fairness. To obtain a design and development of fair AI systems, it can be useful to follow some general principles, which we cite from the White Paper *Artificial Intelligence at the service of the citizen* (Agid 2018). Agid is the public organization that coordinates the activities of the Italian State, its regional and local administration, designing and monitoring the evolution of the IT system of the Public Administration.

Among the recommendations of the White Paper, first of all, what could be defined as a "humanistic" or "anthropocentric" principle should be mentioned, according to which the Artificial Intelligence must always be put at the service of persons and not vice versa. This is a concern, which has always accompanied human beings when inventing new technologies, especially when such tools prove to be able to substitute humans in activities they consider central not only in their everyday life, but even in their definition of what a person is. In this sense, even when it was considered as science fiction, AI has always stimulated the thoughts of many thinkers. For example, Isaac Asimov (Asimov 1950), who wrote his famous laws of robotics: «*a robot may not injure a human being or, through inaction, allow a human being to come to harm; a robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law; a robot must protect its own existence as long as such protection does not conflict with the First or Second Law*». Such criteria can be considered precisely anthropocentric and return to our mind when we read of the concerns about Artificial Intelligence that can "harm" people by deciding or helping to decide whether to give or to deny them a job, a loan, probation, etc.

There are, then, "*principles of equity, such as procedural (non-arbitrariness of procedures), formal (equal treatment for equal individuals or groups) and substantive (effective removal of obstacles of an economic-social nature)*" (Agid, 2018)(p.38). As we have seen, many of the examples we have taken in this article demonstrate that an improperly designed AI may easily violate all these criteria of equity.

More in general, when people is the target of AI decisions, the respect of universal human rights should be the ultimate reference (Noto La Diega, 2018; Raso et al., 2018; Kaye, 2018): along with the technical developments in AI, research communities should spend relevant effort to include a wide range of stakeholders to firstly debate on the understanding of what kind of society we want to build, in order then to better understand how to make the human-AI cooperation work best, without letting the AI rational agents decide in their stead.

5. Conclusions

The advent of Artificial Intelligence will change the way we make use of technology: suffice it to say that already today it can be controlled through speech and that with the passing of the years more and more cognitive tasks will be performed by AI. This is already changing the way we live, and soon certain transformations will become widespread. It is necessary, therefore, to be aware of what we are achieving, in order to collectively design technologies that are both rational and fair. It is not enough, in fact, for AI to be able to carry out its functions driven only by narrow, task-oriented, optimization goals: it is important that it also contributes to building a more just society.

With these premises, following an experience-based line of reasoning, this paper analyzed the limitations of the mainstream Artificial Intelligence definition, which currently drives the development of most AI systems. We picked the example of biases in datasets to show how a too narrow focus on rationality in terms of efficiency and optimization could lead to excessive risks of discrimination towards specific population groups, often disadvantaged. We also gave an overview of other open issues connected to the use of AI agents for decision making purposes -namely liability, transparency, explainability, privacy-, followed by a few general principles for designing AI agents that are more respectful of the humans, principles which were elaborated from the experience of the White Paper *Artificial Intelligence at the service of the citizen* (Agid 2018). We conclude by remarking that only a community effort to study the ethical-social issues hidden behind the mechanisms of design and development of intelligent agents, can lead us to design them in a responsible and inclusive way.

Acnkolewdgments

We would like to thanks Dr. Eleonora Bassi for her precious advises on the legal aspects.

References

Agid - Agenzia per l'Italia Digitale- (2018), *Libro Bianco sull'Intelligenza Artificiale al servizio del cittadino*, Roma, Italy. Available at: https://ia.italia.it/assets/librobianco.pdf

Angwin J, Larson J., Mattu S. and Kirchner L. (2016), "*Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks".* Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Asimov, I. (1950) *I, Robot*. Doubleday, New York City, NY

Barocas, S., Selbst, A.D (2016). "Big data's disparate impact" *104 California Law Reviews, 671, 62 pages*

Bartlett, P.R., Morse, A., Stanton, R.H.,Wallace, N.E. (2017), *Consumer Lending Discrimination in the FinTech Era*, UC Berkeley Public Law Research Paper. Available at SSRN: https://ssrn.com/abstract=3063448 or http://dx.doi.org/10.2139/ssrn.3063448

Bencke M. (2016), "A cautionary tale about humans creating biased AI models". Available at: https://techcrunch.com/2016/09/11/a-cautionary-tale-about-humans-creating-biased-ai-models

Chen, J. (2018), "Fair lending needs explainable models for responsible recommendation", *arXiv preprint arXiv:1809.04684*

Cortez, P. and Silva A. (2008). *Using Data Mining to Predict Secondary School Student Performance.* In: *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) in Porto, Portugal, 2008* . EUROSIS , pp. 5-12, ISBN 978-9077381-39-7. Data available at: https://www.kaggle.com/uciml/student-alcohol-consumption/home

Crawford K. (2016), "Artificial Intelligence's White Guy Problem". Avialable at: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Credit Suisse (2017), "Algorithmic bias: a new fintech challenge", available at: https://qz.com/1121150/algorithmic-bias-a-new-fintech-challenge/ (accessed 30 August 2018)

Di Cosmo R., Zacchiroli S., (2017) "Software Heritage: Why and How to Preserve Software Source Code Inproceedings", *In : Proceedings of 14th International Conference on Digital Preservation (iPRES 2017) in Kyoto, Japan, 2017*, pp. 1-10

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemeln, R. (2012), "Fairness through awareness" *In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ICTS'12) in Cambridge, Massachusetts, 2012*. ACM, pp. 214–226

Floridi, L., Taddeo, M. (2016), "What is data ethics?", *The Royal Society*, Vol. 374 No. 2083

Hardt, M., Price, E. and Srebro, N. (2016) "Equality of opportunity in supervised learning". In: *Proceedings of the 30th International Conference on Neural Information Processing*

Systems in Barcelona, *2016*, Neural Information Processing Systems Foundation Inc., pp. 3323-3331

Kaye, D. (2018), *Report of the Special Rapporteur on the promotion and the protection of the right to freedom of opinion and expression: Note by the Secretary-General,* United Nations, General Assembly. Available at: https://freedex.org/wp-content/blogs.dir/2015/files/2018/10/AI-and-FOE-GA.pdf

Knight W. (2017) , "The Dark Secret at the Heart of AI", available at: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (accessed 30 August 2018)

Larson J., Mattu S., Kirchner L. and Angwin J. (2016), *Compas Analysis*, GitHub, available at https://github.com/propublica/compas-analysis

Lepri B., Staiano J., Sangokoya D., Letouzé E., Oliver N. (2017), "The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good", in Cerquitelli T., Quercia D., Pasquale F. (Ed.s), *Transparent Data Mining for Big and Small Data. Studies in Big Data*, vol 32, Springer International Publishing, Cham, pp. 3–24.

Lichman, M. (2013). *Default of credit card clients Data Set*, UCI Machine Learning Repository, University of California, School of Information and Computer Science (distributor), available at: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Lum, K., Isaac, W. (2016). "To predict and serve?", *Significance 13(5)*, pp. 14–19 .

NY Times Editorial Board (2018), "The Race-Based Mortgage Penalty", availabe at: https://www.nytimes.com/2018/03/07/opinion/mortage-minority-income.html (accessed 30 August 2018)

Noto La Diega, G. (2018) *Against the dehumanisation of decision-making: Algorithmic decisions at the crossroads of intellectual property, data protection, and freedom of information.* Journal of Intellectual Property, Information Technology and Electronic Commerce Law.pp. 1-41

O'Neil, C. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, New York, NY

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., Kim, L. (2018), A*rtificial Intelligence and Human Rights: Opportunities & Risks*, Berkman Klein Center Research Publication No. 2018-6. Available at SSRN: https://ssrn.com/abstract=3259344

Rice, L., Swesnik, D. (2013), "Discriminatory Effects of Credit Scoring on Communities of Color", *Suffolk University Law Review*, Vol. 46, pp. 935-966

Ritter, S., Barrett, D.G., Santoro, A. and Botvinick, M.M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *arXiv preprint arXiv:1706.08606*.

Roy, S. (2018) "Investments in cognitive and AI will reach $19.1b in 2018", available at: https://techwireasia.com/2018/03/investments-cognitive-ai-will-reach-19-1bn-2018/ (accessed 30 August 2018)

Runkin, J. (2018), "Artificial intelligence: €20bn investment call from EU commission " , availabe at: https://www.theguardian.com/technology/2018/apr/25/european-commission-ai-artificial-intelligence (accessed 30 August 2018)

Russell, S.J. and Norvig, P., (2016). *Artificial intelligence: a modern approach*. Pearson Education Limited, Malaysia

Spice, B. (2017), "Questioning the Fairness of Targeting Ads Online", available at: https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html (accessed 30 August 2018)

Sweeney, L. (2013), "Discrimination in Online Ad Delivery", *Queue - Storage*, Vol. 11 No. 3

Yang, T. (2006), "Choice and Fraud in Racial Identification: The Dilemma of Policing Race in Affirmative Action, the Census, and a Color-Blind Society", *Michigan Journal of Race and Law*, Vol. 11 No. 2

Zeileis, A., Meyer, D. and Hornik, K. (2007). "Residual-based shadings for visualizing (conditional) independence." *Journal of Computational and Graphical Statistics*, *16*(3), pp.507-525.

Most relevant policy documents used as input for the Agid White Paper “Artificial Intelligence at the service of the citizen”

Accenture (2018), *Universal principles of data ethics*. Available at: https://www.accenture.com/t20160629T012639Z__w__/us-en/_acnmedia/PDF-24/Accenture-Universal-Principles-Data-Ethics.pdf

Campolo, A., Sanfilippo, M., Whittaker, M., Crawford, K. (2017), AI Now 2017 Report, AI Now Institute, New York, NY. Available at: https://ainowinstitute.org/AI_Now_2017_Report.pdf

IEEE – Advancing Technology for Humanity - (2016), *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems (AI/AS)*. Available at: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf?

Information and Communications Technology Council (2015), *Artificial Intelligence in Canada. Where do we stand?*. Available at: https://www.ictc-ctic.ca/wp-content/uploads/2015/06/AI-White-paper-final-English1.pdf

Inria (2016), *Artificial Intelligence Current challenges and Inria's engagement.* Available at: https://www.inria.fr/en/news/news-from-inria/artificial-intelligence-current-challenges-and-inria-s-engagement

Metcalf, J. Keller, E.F., Boyd, D. (2016), *Perspectives on Big Data, Ethics, and Society*, Council for Big Data, Ethics, and Society. Available at: https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/

National Science and Technology Council - NSTC - (2016), *Preparing for the Future of Artificial Intelligence.* Available at: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., Teller, A. (2016), *Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence*, Stanford University, Stanford, CA. Available at: http://ai100.stanford.edu/2016-report

Authorship details

Author	**Concep-tualization**	**Met hodology**	**Data analysis**	**Data visualization**	**Writing - Draft**	**Writing – Review & Editing**
A. Vetrò	X	X			X	X
A. Santangelo	X				X	X
E. Beretta			X	X		X
J.C. De Martin						X

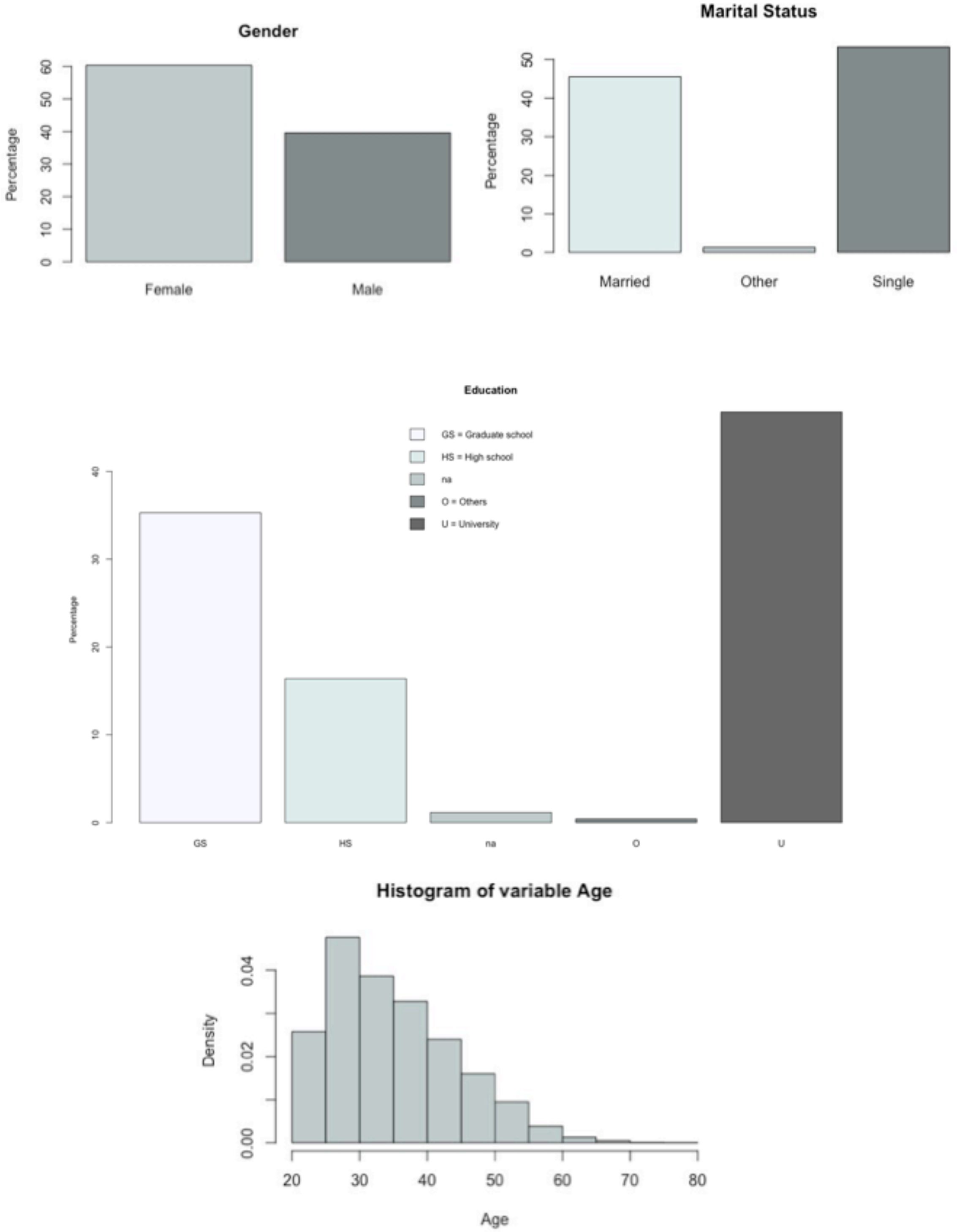

Figure 1 Distribution of selected attributes in the credit card default dataset

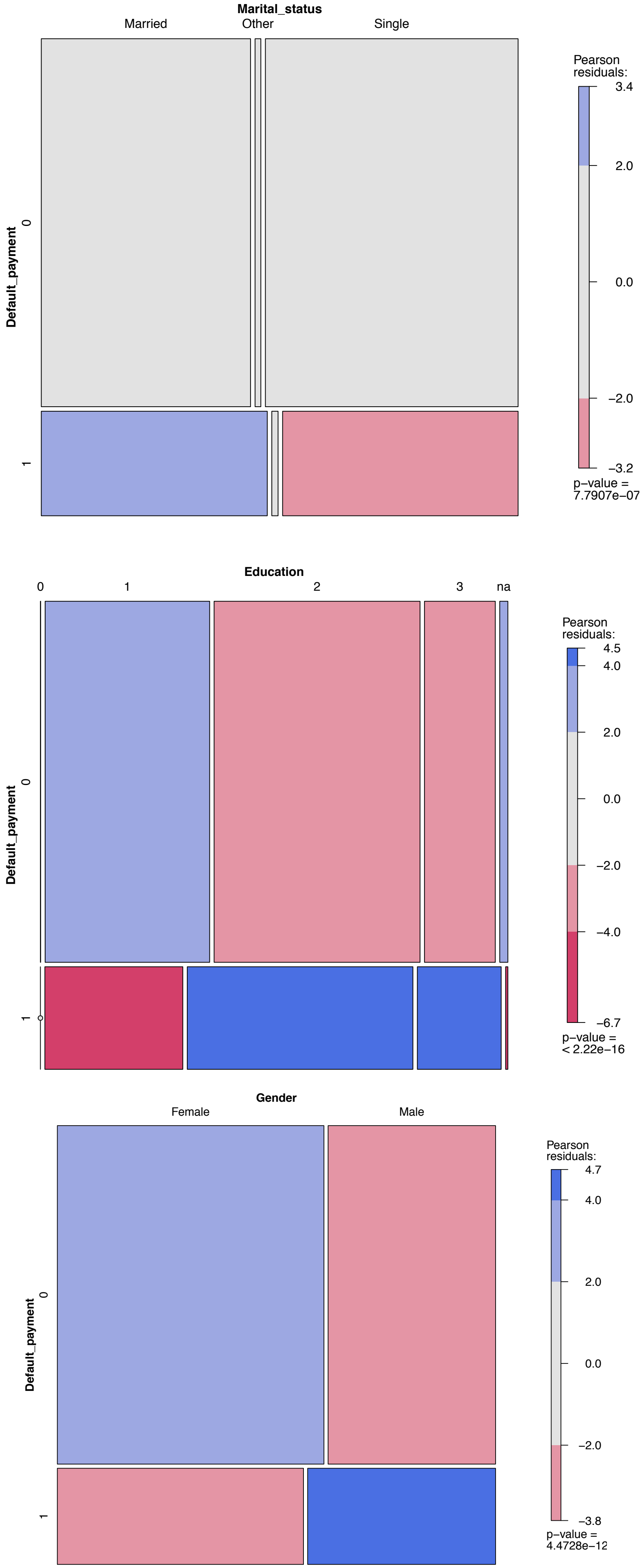

Figure 2 Analysis of disproportions with mosaic plots for selected attributes in credit default dataset