Unveiling the Oracle: Artificial Intelligence for the XXI Century

Federico Cerutti [a]F.Federico Cerutti
Alessia Grassi [b]
A.Alessia Grassi
Mauro Vallati [c]
M.Mauro Vallati

[a] *School of Computer Science and Informatics,*
Cardiff University,
Cardiff, United Kingdom
E-mails: ceruttif@cardiff.ac.uk
[b] *School of Art, Design and Architecture,*
University of Huddersfield,
Huddersfield, United Kingdom
E-mails: alessia.grassi@hud.ac.uk
[c] *School of Computing and Engineering,*
University of Huddersfield,
Huddersfield, United Kingdom
E-mails: m.vallati@hud.ac.uk

The inability of current machines to expose biases induced by programmers and data scientists is leading towards the creation of a new religion, where machines are mystic oracles whose pronouncements have to be believed, and computer users are their servants.

This has to change.

In this paper we discuss the issues that can raise from biases introduced in autonomous systems, with specific care of the case of machine learning systems, and their impact on our society. In the light of the (current and future) exploitation of autonomous systems for law enforcement and war-fighting, we emphasise the importance of issues related to discrimination and safety. We also support the bold claim that artificial intelligence can help artificial intelligence in overcoming those issues: by enabling artificial intelligence to record every single step that lead to a given inference, and to argue with humans, we can unveil the mystic oracle and trust its services.

Keywords: Artificial Intelligence, Argumentation, AI and Society

Introduction

Indisputably, Artificial Intelligence (AI) is already strongly affecting our lives and our society. AI approaches are exploited in a wide range of applications, such as: for understanding and driving consumer behaviour,[1] for supporting decision-makers and the decision-making process, for fostering creativity,[2] or for identifying suspicious behaviours. Despite the widespread presence of AI in our society, it is well-known that, thanks to advances in computational power and in to the optimisation of the approaches, the exploitation of AI-based techniques is steeply rising. However, machines—as we learnt to know them in these early years of the 21st century—are as biased as their programmers [6]. Moreover, differently from (some) humans, machines lack the ability to expose and discuss their biases. This is raising significant concerns on the impact of machines on human activities[3] and society organisation,[4] to mention a few. That is because, arguably, using a machine without knowing in the minimal details the way it has been programmed, is similar to ask a mystic oracle whose pronouncements have to be believed by faith [23].

This has to change.

Research has already been carried out for unveiling part of such a mystic oracle, thanks also to events like the Workshop on Human Interpretability in Machine Learning[5] or the Workshop on Explainable Arti-

[1] `http://www.independent.co.uk/life-style/christmas/bot-until-you-drop-a7477406.html` (on 29/05/2017)

[2] `https://singularityhub.com/2017/02/06/art-in-the-age-of-ai-how-tech-is-redefining-our-creativity/` (on 29/05/2017)

[3] `https://www.theguardian.com/technology/2017/feb/09/robots-taking-white-collar-jobs` (on 29/05/2017)

[4] `https://www.theguardian.com/technology/2017/mar/13/artificial-intelligence-ai-abuses-fascism-donald-trump` (on 29/05/2017)

[5] `https://sites.google.com/site/2016whi/` (on 29/05/2017)

AI Communications
ISSN 0921-7126, IOS Press. All rights reserved

ficial Intelligence.[6] Remarkable contributes from such events include the work of Lipton [19], which provides a refined definition of *interpretability* of predictive models, and the work from Krause et al. [17] that introduces visual techniques for interpreting some type of predictive models. On a similar line of research, Letham et al. [18] presented a methodology for deriving accurate predictive models that also human experts can interpret. Their approach lies on decision lists, which consist of a series of if-then statements that transform high-dimensional, multivariate feature spaces into a series of simple, readily decision statements.

Research like [18,17] is clearly more than welcome, but it does not solve all the issues associated to using such a mystic oracle. As we will develop further in Section , the mystic oracle relies on various components, each of those plays a specific role, and may very well be affected by human biases or mischievous actions. For example, what would happen if an autonomous system for detecting suspicious activities, that might lead to searches and seizures by the police, is trained on a specific dataset where most of the training cases labelled as *suspicious activity* show a male, black individual? Would the autonomous system infer that the gender and the skin colour are relevant features? And what would happen if such a system becomes widely used in policing actions? Under the UK law, every individual can be stopped and searched if policemen have "reasonable grounds" for suspicious activities.[7] In the USA, the fourth amendment provides "[t]he right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause."[8] Would such an autonomous system be used for providing "reasonable grounds" or "probable cause" for searches, or seizures? We, the researchers in Artificial Intelligence, have the moral obligation to educate the society on the issues that may raise when using autonomous systems, and in Section we explore some of those issues.

Indeed, the goal of this paper is twofold: on the one hand, we want to raise the awareness of these issues within the ACM community and with the general public. On the other hand, we aim to discuss a reasonable, methodological pathway for addressing these issues. Our belief is that artificial intelligence can help artificial intelligence to overcome (some of) those issues, and we discuss it at length in Section . Finally, we conclude the paper with general discussions and conclusion in Section .

The Mystic Oracle

Diakopoulos [7] describes a generic autonomous agent on the basis of five main components: *human involvement* and accountability; *data* used as input; *model* for manipulating the input; *inferencing* and issues associated to accuracy; and *algorithmic presence* as the impact that might have on the ultimate user knowing that the given inference has been curated by an autonomous agent. Although these elements are relevant for many types of AI approaches, in this paper we will consider mostly the case of machine learning systems because of their massive diffusion and exploitation.

Formally, a computational system is said to learn from experience E, with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [21].

Practically, machine learning approaches generate predictive models on the basis of experience gained by analysing a provided training data set. Training instances are described through some available observed characteristics (features). The training set is, by assumption, representative of the population on which the trained model will be exploited. There exists a wide range of techniques for generating predictive models—such decision trees, neural networks, support vector machines, etc.—but they all try to identify some sort of patterns in the values of the features of the given training examples, that allows to provide accurate predictions. It shall be noted that the type of features (e.g., continuous values, classes, etc.), as well as the type of prediction, has a strong impact on the algorithms for generating predictive models.

Figure 1 depicts the above mentioned five components introduced by Diakopoulos [7], and the interactions among them that we will consider in this paper. It is beyond doubt that, in machine learning approaches, a model is derived from data, and leads to a series of inferences. Moreover, human is usually –but not always– involved in the process of selecting relevant data/for-

[6] http://home.earthlink.net/~dwaha/research/meetings/ijcai17-xai/ (on 29/05/2017)

[7] http://www.legislation.gov.uk/ukpga/1984/60/contents (on 29/05/2017)

[8] https://catalog.archives.gov/id/1408042 (on 29/05/2017)

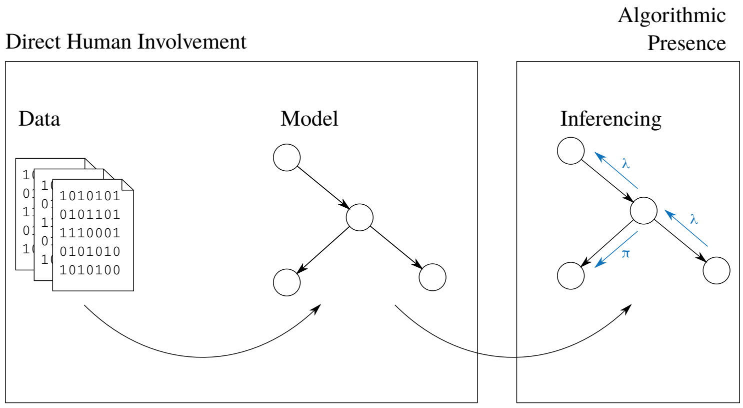

Fig. 1. Fundamental components of an autonomous agent: a kite-level picture.

malism to be used, as well as the general model to be used in a specific scenario, e.g. Bayesian network vs neural network, etc. Finally, algorithmic presence is clearly determined by the use or not of a given model.

The Veiled Oracle

With reference to Figure 1, we argue that biases and mischievous actions can affect each of the five components. However, given the venue for this paper, we will abstain from discussing how biases can be originated in humans, therefore we will look more closely at the selection of *data*, the choice of *model*, the quality of *inferencing*, and at the effect of *algorithmic presence*.

Data and Model

First of all, it is worth remember here that each autonomous system "is limited in making its predictions to analysis of the data within its dataset, and it cannot consider other facts that might be relevant but that were not included. In contrast, human beings are always at least potentially capable of including a new piece of relevant information in an analysis" [29].

Bearing in mind such caveat, Citron and Pasquale [6] discuss predictive algorithms assessing risks, desirable employees, reliable tenants, etc. However, they specifically highlight how datasets may contain "inaccurate and biased information." Indeed, as noted by van Eijk [35], criminologists and legal scholars have criticised risk assessment tools for their potential bias against racial/ethnic minorities and women, which could result in sentencing disparities, e.g. [14,4,32]. As noticed by Skeem and Lowenkamp [32], criminal history mediates the relationship between race and future arrest.

Those are evidence of possible *implicit discrimination*, which can be of three types: *masking*, *subconscious discriminatory motivations*, and *relying upon tainted datasets or tools* [42]. The first type collects all mischievous activities aimed at making discriminatory practices in a way in which discrimination might prove undetectable or at least defensible [2].

With regards to the second type, Citron and Pasquale [6] highlight how "human beings [. . .] biases and values are embedded into the [. . .] predictive algorithms," thus giving rise to subconscious discrimination. For instance, as reported by Friedmanand Nissenbaum [10], the National Resident Match Program (NRMP)[9] implements a centralised method for assigning medical school graduates their first employment following graduation in USA. The original algorithm was designed at a time where there were very few couples—both partners graduates from medical school—involved. In late 1970s, more women entered medical schools and thus more married couples sought medical appointments though the NRMP. At this point, it was discovered that the original admissions algorithm placed married couples at a disadvantage in achieving an optimal placement, as compared with their single peers [30].

Finally, Zarksy [42] discusses at length how discriminatory outcomes are generated by "recklessly—or perhaps, mere negligently—relying upon tainted datasets." It is beyond the scope of this paper to discuss the subtle distinction between reckless and negligence. However, this is perhaps the simplest case of bias a computer or data scientist encounters. Is the dataset

[9] http://www.nrmp.org/ (on 29/05/2017)

representative enough? Are we applying a model violating the assumptions behind its training?

This naturally raises questions about accountability of the *human involvement*. Discussions around this theme now dominate chronicles, and examples include autonomous cars, and more dangerously autonomous weapons. On 28th July 2015, at the opening of the International Joint Conference on Artificial Intelligence in Buenos Aires, Argentina, an open letter was published[10] to join the Human Right Watch[11] in asking for a ban on the development of autonomous weapons. As stated in the Human Right Watch's document, "the use of fully autonomous weapons raises serious questions of accountability, which would erode another established tool for civilian protection. Given that such a robot could identify a target and launch an attack on its own power, it is unclear who should be held responsible for any unlawful actions it commits."

Inferencing and Algorithmic Presence

When faced with the result of an algorithmic procedure, people often experience "automation bias," which is "the tendency to trust an automated system, in spite of evidence that the system is unreliable, or wrong in a particular case" [1]. Moreover, Kirkpatrick [16] discusses legal implications that can arise when an autonomous system is used in risk assessment of a suspect. In July 2016 the Wisconsin Supreme Court sentenced that warnings must be attached to the scores to flag such a system's "limitation and cautions".[12]

The way Rich [29] analyses the result of an inference produced by an autonomous agent is to consider it at the same level of an informant, in that they are outside of a law enforcement and provide information about criminal activity. Informants can be divided in: (1) criminal informants; (2) anonymous tipsters; and (3) citizen-informants [39]. Criminal informants generally ask in exchange money or leniency. Anonymous tipsters do not disclose identifying information. Citizen-informants provide information by virtue of being the victim of or witness to a crime.

Rich [29] argues that an autonomous system would not seat in any of these three traditional categories. In one sense, autonomous systems' designers are like citizen-informants. However, like anonymous informants, the designers almost certainly will not be subject to criminal prosecution if a prediction is wrong. Moreover, a designer is motivated (at least in part) by money.

Finally, even if a legal framework for handling information coming from an autonomous system was available, a more serious issue still stands, namely the difficulty that most humans have to fully understand uncertainty. For instance, Gigerenzer [13] discusses at length the fact that even physicians are often confused about the notions of sensitivity and specificity, and their effect on conditional probabilities. Although no systematic studies of effects on patients exists, there are anecdotal reports of people with HIV false-positive test results who: engaged in unprotected sex with HIV-positive people believing that it would not matter anymore; committed suicide; and endured harmful effects of unnecessary antiretroviral treatments [12].

Unveiling the Oracle

Our main claim is that most of—if not all—the issues highlighted in Section can be addressed building on top of existing artificial intelligence techniques. At the same time, it is beyond the scope of this limited paper to show complete solutions to those issues. Rather, we choose to consider a sub-field in artificial intelligence, namely argumentation theory, and look at how it would help in addressing those issues.

Argumentation theory sits in the intersection of three major lines of research: informal logic, philosophy, and non-monotonic reasoning. Stephen Toulmin in his book The Uses of Argument criticised traditional studies in deductive reasoning and logic. This lead him to introduce a *warrant* for each line of reasoning allowing for rebutting such a line of reasoning when specific circumstances happened thus affecting the warranty. Other works on rebuttals were influential for John Pollock, who migrated from philosophical studies to artificial intelligence in order to formalise the intuitive concept of *defeasible reasoning*, now at the hearth of argumentation theory. Finally, argumentation theory borrows largely from the field of non-monotonic reasoning, especially for what concerns reasoning systems: the first systems were heavily based on logic programming.

The field is now start testing the claim that argumentation theory "captures naturally the way humans argue

[10] `https://futureoflife.org/open-letter-autonomous-weapons/` (on 29/05/2017)

[11] `https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots` (on 29/05/2017)

[12] `https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690` (on 29/05/2017)

to justify their solutions to many social problems" [8]. Recentely, an argumentation-based system has been used to aggregating clinical evidence on lung cancer treatments [40], taking into account subjective criteria such as preferences over outcome indicators. Moreover, experiments with human participants are carried out, and projects are proposed to apply argumentation theories to a large variety of aspects related to human cognition. For instance, Framework for Computational Persuasion[13] is a project under development to build a framework for computational persuasion to support behaviour change using knowledge representation and reasoning techniques, together with computational models of argument, to offer an argument-based approach to persuasion.

Background in Argumentation Theory

Since the oracle we want to unveil is, unquestionably, a logical one, let us focus on logical aspects of argumentation brilliantly summarised by Chesñevar et al. [5] in their survey paper to which we refer interested readers.

Formal logic of arguments emerged in the last 30 years as one style of formalising non-monotonic reasoning which differentiates from classical logic for its ability to deal with incomplete and potentially inconsistent information [20]. Among other founding fathers of argumentation theory, Pollock [27] moved his theory of defeasibily—originally used within epistemology to address questions of justification—into computer science, developing a theory of defeasible reasoning. In his view [27], reasoning proceeds by constructing arguments based on *reasons* that can be either *conclusive*—i.e. non defeasible—or *prima facie*—i.e. that create a presumption in favour of their conclusion but they can be defeated. Pollock was a strenuous supporter of statistical generalisation and thus, there should be little surprise in reading that "a rational agent must be equipped with rules (1) enabling it to form beliefs in statistical generalisations, and (2) enabling it to make inferences from those statistical generalization to beliefs about individual matters of fact" [27, p. 59].

Pollock's work was very influential for a large part of the argumentation community, who started building rule-based systems incorporating the distinction between *defeasible* and *strict*, i.e. non defeasible, rules, e.g. [37] among others. Although a logical language is often used to express such rules, Walton et al. [38] considered the more informal concept of *argumentation scheme* as a reasoning pattern aimed at supporting the acceptance of a conclusion. This line of research benefited from Dung's theory [8] that abstracts from the inner structure of arguments focusing only on defeaters (or attackers) for deriving criteria to determine set of arguments that are collectively acceptable, i.e. that collectively *survive* the attacks they receive. In this context, criteria are referred to as semantics.

Argumentation and Data

From our perspective, issues highlighted in Section with regards to data belong to two classes that can be summarised by two questions: *are we using the right information*? And, *are we using quality data*?

As per the first question, the issue is far from being settled. In more computational terms it can be rephrased as: what are the criteria of adequacy of logical formalisation of natural language statements? [26]. In this respect, there are two main schools of thoughts, one promoting *atomism*, the other *reflective equilibrium*. The former one, advocated for instance in [3], suggests that the adequacy of a formalisation of a sentence does not depend on formalisation of other sentences. The latter, advocated for instance in [26], argues that formalisation is inherently a holistic endeavour. We believe that this question should be settled for the sake of artificial intelligence, leaving aside benefits for the philosophy of logic. Having clear criteria to use for testing the quality of formalisation and thus answering the question *are we using the right information?* is of primary importance.

The second question—i.e. are we using quality data—highlights the need of confidence in evidence-based reasoning.[14] To this end, Fox [9] contends that a theory of arguing about evidence should have ten properties:

1. it should follow an argumentation process constructing reasons for/against competing claims;
2. evidential arguments should increase/reduce confidence in claims;
3. *ceteris paribus*, the more independent and sound arguments for a given claim, the greater our confidence in such a claim;

[13] `http://www.computationalpersuasion.com/` (on 7th May 2017)

[14] Unless *alternative facts* `https://www.youtube.com/watch?v=VSrEEDQgFc8` (on 29/05/2017) are preferred to ground inferencing

4. a single argument can be conclusive for confirming or refuting a claim;
5. arguments and theories can themselves be questioned;
6. some arguments can be stronger than others;
7. in the absence of information about relative strength, contradictory arguments still play an important role in decision making;
8. it is desirable to develop systems using sound, formal languages for argumentation but that can be translated to and from intuitive natural language interfaces;
9. a rational agent can choose the hypothesis that has the greatest confidence among all the competing hypotheses, unless there are grounds to argue against such a confidence;
10. a rational agent not forced to choose may defer a decision on the grounds that the arguments are unwarranted.

To this end, the CISpaces tool [34] is, to our knowledge, the most advanced attempt to satisfy most of those properties. It supports intelligence analysts in their sense-making activities exploiting argumentation schemes [38] for building arguments, thus satisfying the first property, as well as supporting critical thinking. Moreover, CISpaces allows analysts to inspect the provenance of data, recorded using the PROV-DM model[15] thus supporting strength—in the form of preferences—in favour of different claims based on evidence (second, third, and sixth properties) and ensuring accountability. Moreover, it builds on top of ASPIC+ [28] and uses Dung's preferred semantics [8], and thus satisfies the fourth and seventh properties. CISpaces does not allow to express arguments against the argumentation schemes used—although we are not aware of approaches in that direction that seems to us borderline with the first question above, i.e. *are we using the right information?*—and it does not have a natural language interface. Regarding the last two properties, CISpaces is not aimed at replacing the human decision maker, rather to support them, therefore the burden of the decision is left always on the shoulders of the rational agent who can choose the hypothesis that has greatest confidence, or defer the decision, thus suggesting that more information might be needed.

Listing 1: An example of Bayesian Rule Lists from [18, Fig. 3]. In parentheses is the 95% credible interval

```
if hemiplegia and age > 60
     stroke risk 58.9% (53.8%−63.8%)

else if cerebrovascular disorder
     stroke risk 47.8% (44.8%−50.7%)

else if ...
```

Argumentation and Model

Letham et al. [18] present a machine learning system where the inferencing step is based on a decision list that is mined from data. This reflect, in spirit, [15], where Holte argues that there are cases in which the accuracy of rules that classify examples on the basis of a single attribute is sometime higher than other more complex approaches. Letham et al. [18] show that their Bayesian Rule Lists (BRL) aims at hitting "the *sweet spot* between predictive accuracy, intrepretability, and tractability" by providing (1) more accurate results than SVM, Random forests, and others systems, on the task of predicting stroke risk on real data; and (2) and interpretable model of the form of the one in Listing 1.

Although Listing 1 already provides a fairly understandable prediction model, there are two main issues that need to be addressed. The first one is to explain the meaning of those percentages, and to this aim we believe that the discussion on provenance we had in Section is very relevant in this context as well. For instance, knowing the details of the chosen training dataset would be extremely relevant to understand whether the prediction can actually apply to the given case.

The second issue is how an expert can interact with the system in order, perhaps, to correct the mined model or to ask questions such as "what if. . . ." This should not be considered as a minor detail. Indeed, Citron and Pasquale [6] highlighted this as one of the main requests in medical and legal domain—a requirement if you want, given that barristers and physicians are among the most important stakeholders of data mining.

While we are not aware of any approach in argumentation aimed at addressing what-if analysis—an

[15] http://www.w3.org/TR/prov-dm/ (on 29/05/2017)

area that we believe should be definitely subject of future studies—argumentation community addressed other elements of interaction with machine learning systems. Notably, Možina et al. [22] combine machine learning with experts' arguments or reasons for some of the learning examples thus constraining the combinatorial search among possible hypotheses. We believe that it would be very beneficial to further investigate connections with [18]. Moreover, in the context of the RoboCup, Gao and Toni [11] present a way for experts to provide knowledge in the form of heuristics to a system for reinforcement learning. It would also be clearly very beneficial to investigate further how this might link in the context of active learning [31]—i.e. where an active learner may question a human expert that already understands the nature of the problem.

Argumentation and Inferencing

To exemplify an inferencing case, let us consider the case that a Bayesian network—a directed acyclic graph where nodes represents stochastic variables, and arrows dependencies that can be specified by the means of conditional probabilities—has been used by an autonomous agent as a model to make predictions. That is, the posterior probability of a specific variable in a Bayesian network is derived on the basis of evidence. The question is how such an inferencing step can be explained to a human agent.

In a recent paper, Timmer et al. [33] discuss an approach to derive explanatory arguments from a Bayesian network. Pivotal in their work is the notion of a *support graph* constructed for a variable of interest that captures the support the variable of interest receives from the other variables. Then, once evidence is provided, the support graph is used to derive arguments that describe the *logical steps* needed to interpret the Bayesian judgement of the variable of interest.

Despite Timmer et al. [33] seem not to be aware of [18], it is almost immediate to draw parallels between the two approaches: the Bayesian Rule List derived using [18] seems to represent a very specific type of support graph. The main difference is that in [18] the inferencing step is provided by the Bayesian Rule List itself, rather than an underlying Bayesian network.

However, both [33] and [18] suffer from the lack of empirical evaluation of the quality of their explanations with human judges. This has been the subject of [25], where it is shown that Causal Explanation Tree [24] and Most Relevant Explanation [41] models provide better fits to human data. This is encouraging especially because, intuitively, Causal Explanation Trees seem expressive enough to capture Bayesian Rule Lists.

Argumentation and Algorithmic Presence

In Section we reported that Rich [29] would consider autonomous agents at the same level of informants. Assuming that an autonomous agent can also provide a confidence interval (cf. for instance Listing 1), then this should be the case when the confidence in the answer is *low*.[16] However, there are situations where the inferences produced by an autonomous agent are significantly accurate, and thus they would appear more like an expert opinion. However, while human experts are often allowed "to draw on their own experience and specialised training to make inferences from and deductions about the cumulative information available to them that *might well elude an untrained person*"[17] without the need to prove the soundness of the reasoning behind their inferences and deductions, this should not be allowed for an autonomous agent.

For instance, we can build on top of Walton et al. [38] schemes for argument from witness testimony and from expert opinion to derive the following scheme for evidencing and questioning the algorithmic manipulation of a piece of information we received:

Argument from Autonomous Inferencing

Major Premise: A is an autonomous system trained in subject domain S containing proposition P.
Minor Premise: A asserts that proposition P is true (or false).
Conclusion: P is true (or false).

Critical questions

CQ1: What are A's maker interests?
CQ2: Is A's assertion internally consistent?
CQ3: Is A training adequate to make a judgement about P?
CQ4: Is the provenance of A's judgement about P sound?
CQ5: Is A's assertion consistent with the known fact of the case (based on evidence independent from A)?

[16] How to label a numeric value as *low* is not trivial. A confidence interval of 20%–30% might be *low* if we are contemplating the risk of rain in London; but it might be very *high* if it is the risk of a terrorist attack in New York

[17] `https://www.law.cornell.edu/supremecourt/text/534/266` (on 29/05/2017)

CQ6: Is A's assertion consistent with other, independent autonomous systems' assertions?

In particular, to minimise the "automation bias" highlighted by Asaro [1] and to ensure accountability, we strongly believe that no information provided by an autonomous system should be accepted until all the critical questions have been successfully addressed. This will then require to record: eventual conflict of interest with the individual of company that created the autonomous system; the internal consistency of the system; the dataset(s) used for training; all the steps that went from the data collection of the dataset, to the generation of the model, to the collection of data specifically used for prediction, to the inferencing step. Finally, we believe it should be enforced the general principle that every piece of information requires confirmation from multiple independent sources before being accepted.

Discussion

The great challenge of Artificial Intelligence is "to understand the nature of intelligence and cognition so well that computers can be made to exhibit human-like abilities" [36, Preface].

We argue that Artificial Intelligence can go beyond enabling computer to exhibit *everyman-like abilities* such as driving a car: it can enable computers to collaborate with us at the highest scientific standards. In fact, research is currently being carried out for investigating how AI can be exploited for *evolving* some of the most pivotal human society processes, such as the democratic process,[18] and for addressing difficult social problems.[19]

To successfully achieve that, we believe an holistic, homogeneous approach that addresses all the components depicted in Figure 1 of an autonomous system is not only desirable: it is necessary. Several pieces of research already addressed many of those components separately: it is now necessary to connect the dots, to draw the big picture, and to evaluate it by the means of extensive experimentation with a large variety of human subjects.

There lies before us, if we choose, a continual progress and benefit in this fourth industrial revolution with true human-machine collaborations. Otherwise, we risk the rise of a new religion where machines are mystic, veiled, ubiquitous oracles, and their pronouncements and decisions will have just to be accepted. Until a new revolution.

[18] https://linkeddemocracy.com/ (on 29/05/2017)

[19] http://cais.usc.edu/ (on 29/05/2017)

References

[1] P. M. Asaro. Modeling the moral user. *IEEE Technology and Society Magazine*, 28(1):20–24, 2009.

[2] S. Barocas and A. Selbst. Big Data's Disparate Impact, 2016.

[3] M. Baumgartner and T. Lampert. Adequate formalization. *Synthese*, 164(1):93–115, aug 2007.

[4] J. L. Chenane, P. K. Brennan, B. Steiner, and J. M. Ellison. Racial and Ethnic Differences in the Predictive Validity of the Level of Service Inventory–Revised Among Prison Inmates. *Criminal Justice and Behavior*, 42(3):286–303, 2015.

[5] C. I. Chesñevar, A. G. Maguitman, and R. P. Loui. Logical models of argument. *ACM Computing Surveys (CSUR)*, 32(4):337–383, 2000.

[6] D. K. Citron and F. A. Pasquale. The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89:1–33, 2014.

[7] N. Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.

[8] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[9] J. Fox. Arguing about the evidence: a logical approach. In *Proceedings of the British Academy*, volume 171, pages 151–182, 2011.

[10] B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347, 1996.

[11] Y. Gao and F. Toni. Argumentation accelerated reinforcement learning for cooperative multi-agent systems. In *Proc. of ECAI*, pages 333–338, 2014.

[12] G. Gigerenzer. *Reckoning with risk: learning to live with uncertainty*. Penguin UK, 2003.

[13] G. Gigerenzer. *Simply rational: Decision making in the real world*. Oxford University Press, USA, 2015.

[14] K. Hannah-Moffat. A conceptual kaleidoscope: contemplating dynamic structural risk' and an uncoupling of risk from need. *Psychology, Crime & Law*, 22(1-2):33–46, 2016.

[15] R. C. Holte. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, 11(1):63–90, 1993.

[16] K. Kirkpatrick. It's Not the Algorithm, It's the Data. *Commun. ACM*, 60(2):21–23, 2017.

[17] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proc. of CHI*, pages 5686–5697, 2016.

[18] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, sep 2015.

[19] Z. C. Lipton. The mythos of model interpretability. In *The Workshop on Human Interpretability in Machine Learning*, 2016.

[20] J. McCarthy. Circumscription–A form of non-monotonic reasoning. *Artificial Intelligence (Special Issue on Non-Monotonic Logic)*, 13(1-2):27–39, 1980.

[21] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., 1997.

[22] M. Možina, J. Žabkar, and I. Bratko. Argument based machine learning. *Artificial Intelligence*, 171(10-15):922–937, jul 2007.

[23] Nature. Digital intuition. *Nature*, 529(7587):437, jan 2016.

[24] U. H. Nielsen, J.-P. Pellet, and A. Elisseeff. Explanation trees for causal Bayesian networks. In *Proc of UAI*, pages 427–434, 2008.

[25] M. Pacer, J. Williams, X. Chen, T. Lombrozo, and T. L. Griffiths. Evaluating Computational Models of Explanation Using Human Judgments. In *Proc. of UAI*, pages 498–507, Arlington, Virginia, United States, 2013.

[26] J. Peregrin and V. Svoboda. Criteria for logical formalization. *Synthese*, 190(14):2897–2924, apr 2012.

[27] J. L. Pollock. *Cognitive Carpentry. A Blueprint for How to Build a Person*. MIT Press, Cambridge, MA, 1995.

[28] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument* {\&} *Computation*, 1(2):93–124, jun 2010.

[29] M. Rich. Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment, 2016.

[30] A. E. Roth. The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory. *Journal of Political Economy*, 92(6):991–1016, dec 1984.

[31] B. Settles. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6:1–114, 2012.

[32] J. L. Skeem and C. T. Lowenkamp. Risk, Race, and Recidivism: Predictive Bias and Disparate Impact. *Criminology*, 54(4):680–712, 2016.

[33] S. T. Timmer, J.-J. C. Meyer, H. Prakken, S. Renooij, and B. Verheij. A two-phase method for extracting explanatory arguments from Bayesian networks. *International Journal of Approximate Reasoning*, 80(C):475–494, 2017.

[34] A. Toniolo, T. J. Norman, A. Etuk, F. Cerutti, R. W. Ouyang, M. Srivastava, N. Oren, T. Dropps, J. A. Allen, and P. Sullivan. Agent Support to Reasoning with Different Types of Evidence in Intelligence Analysis. In *Proc. of AAMAS*, pages 781—-789, 2015.

[35] G. van Eijk. Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society*, 2016.

[36] F. van Harmelen, V. Lifschitz, and B. Porter. *Handbook of Knowledge Representation*. Elsevier Science, San Diego, USA, 2007.

[37] G. A. Vreeswijk. *Studies in Defeasible Argumentation*. PhD thesis, Department of Computer Science, Free University Amsterdam, 1993.

[38] D. Walton, C. Reed, and F. Macagno. *Argumentation schemes*. Cambridge University Press, NY, 2008.

[39] A. C. Werner. What's in a Name? Challenging the Citizen-Informant Doctrine. *New York University Law Review*, 89(6):2336–2380, 2014.

[40] M. Williams, Z. W. Liu, A. Hunter, and F. Macbeth. An updated systematic review of lung chemo-radiotherapy using a new evidence aggregation method. *Lung cancer (Amsterdam, Netherlands)*, 87(3):290–5, mar 2015.

[41] C. Yuan, H. Lim, and T. Lu. Most Relevant Explanation in Bayesian Networks. *Journal of Artificial Intelligence Research*, 42:309–352, 2011.

[42] T. Zarsky. Understanding Discrimination in the Scored Society. *Washington Law Review*, 89(4):1375–1412, 2015.