Artificial Intelligence and Human Dialogue

Silvia Crafa

Dipartimento di Matematica "Tullio Levi-Civita"
Università di Padova
silvia.crafa@unipd.it

Abstract: The complexity of the impact of Artificial Intelligence (AI) systems on society and on people's lives calls for a dialogue between people with different expertises and different roles in society. In this article we discuss many facets of this dialogue, exploring specific issues that emerge in AI cross-disciplinary research, and showing how different approaches and methodologies can cooperate and produce new insights without losing their specificity. We also remark the significance of a dialogue between science and technology, and that between the research community and the society, which bring to light the different responsibilities involved in an ethical approach to artificial intelligence.

Keywords: *Artificial Intelligence, Cross-disciplinary research, Research Languages and Methodologies*

1. Introduction

Artificial Intelligence is a highly overloaded term: even as a computer science research field, a precise definition of artificial intelligence is missing. It ranges from any software system performing complex tasks such as controlling the flight of an airplane, through a broad range of machine learning algorithms, to the embodied digital technologies of robotics. Generally speaking, the distinctive characteristic of AI systems is that they feature complex cognitive tasks, that is, they exhibit an advanced processing of inputs that can be compared to human understanding of images or voice. The surprising classification and prediction performances of learning algorithms then suggest the idea that the machine really learned and possibly understood something, while it actually just optimized a (huge) number of parameters searching into a given (rich) set of solutions. Moreover, differently from other complex software systems, but similarly to human understanding, AI's data processing is able to flexibly adapt, at some extent, to environmental changes, so that its behaviour can be tailored to the user or agent interacting with the machine.

On the other hand, the term Artificial Intelligence has now moved beyond computer science and digital technologies to enter many other fields, like economy, politics, law, philosophy, social sciences, neuroscience, psychology, education, and it is a recurring theme also in mass media and in the public discourse. In each of these domains the term is enriched with a lot of specific connotations which often remain implicit, even if sometimes the implications of such hidden meanings or nuances can become very explicit, resulting for instance in an official regulation or in very focused public and private investments[1].

In such a context, a fundamental role can be played by cross-disciplinary research, which is mainly based on human dialog. First, since AI-based tools and services will have an impact on many aspects of human life, their investigation and development cannot be committed only to scientists and engineers; experts of social, cognitive, ethical and legal issues should be involved since the early stages of the design level and not just in a post hoc assessment. Secondly, the difficult process of establishing a cross-disciplinary dialogue based on a common language usually uncovers implicit assumptions and gives new insights about a hypothesis or a definition. Indeed, nuances and hidden connotations are naturally attached to terms and concepts that are routinely used in specific research communities, forming a sort of jargon. Such an implicit knowledge comes to light when talking with people with a

[1] Powles 2017; Harwell 2018; European Parliament 2017; UNHR 2018

different background and trying to transfer a term from a jargon to another. In Section 2 we will discuss this issue, presenting a number of problematic examples that illustrate how a cross-disciplinary dialogue not only may advance the general knowledge, but also produces deeper insights within each single discipline.

Another kind of human dialogue that is worth enhancing is the one between science and technology, which are particularly tangled in the context of AI. We will discuss it in Section 3, observing that such a dialogue brings in the relationship between theory, knowledge and applications, in terms of both feasibility and impact on the society. Therefore, the importance of an ethical framework guiding both scientific and technological developments is remarked. Finally, Section 4 broadens the dialogue beyond experts, concentrating on the mutual benefits coming from a dialogue between the research community and the society. The observation that modern digital systems are actually socio-technical systems entails researchers' social and ethical responsibilities and at the same time calls for an educated social conversation that could effectively prompt and orient the priorities of research investigations and practical applications.

2. The Cross-Disciplinary dialogue

A main distinction between academic disciplines is between scientific disciplines and humanities. Besides the topics of interest, they differ in their research methodology: while the scientific method is grounded in a manipulative experimental method and a rigorous formalization based on maths, the humanities acquire new knowledge by means of critical, speculative and comparative methods. As AI enters areas like medicine, criminal justice, labour, and financial markets, a conversation between different expertises and an appropriate combination of different research methods are of paramount importance to cross-fertilize ideas, provide insights and prevent failures.

In this section we focus on the problem of combining investigation methods coming from different disciplines, exploring a number of issues in the specific context of AI research.

2.1. To define or not to define?

The scientific approach of (computer) scientists requires that the development of a theory or of an artifact (an algorithm, a software, a tool, a machine) be based on well-founded *definitions*. This approach ensures that, independently of the names chosen by researchers to identify a concept, its meaning and the results that logically and mathematically derive from it can always be disambiguated by looking at its definition. The intent of a formal definition is also to remove connotations and alternative meanings that are usually attached to a term of a human language. It is a philosophical question to understand to what extent formal, i.e., mathematical and logical, definitions have the intended univocal meaning. On the other hand, there are subjects that by their very nature have a complex and broad meaning; for instance, the notion of health, justice, due process, discrimination, ethics and also intelligence, can hardly be cut down to a mathematical definition or to a metric.

Here a main difference between human and artificial intelligence emerges. *Ambiguity* is a constitutive element of human intelligence: people effectively, and creatively, engage in many social activities by relying on ambiguous terms and concepts, whose semantics is "defined" pragmatically during their actions. Machines require instead explicit representations, such as ontologies or semantic networks, to deal with imprecise concepts. Interestingly, the *reinforcement learning*[2] AI technique can be seen as using a sort of pragmatic semantics. In this case, the autonomous agent evaluates and interacts with the environment in which it operates, reacting to environment changes (possibly caused by its previous actions) so to maximize a given goal. In particular, the agent receives a reward or a penalty for the re-action it performed, so to calibrate its learning algorithm and choose the suitable next action. Reinforcement learning is an effective solution in practical applications where the agent's environment is not completely defined, hence somehow ambiguous (model-free technique[3]), however it is far from an "intelligence" working with ambiguous concepts.

Therefore, a useful outcome of a cross-disciplinary dialogue would be the identification in AI techniques of both ambiguities that should be better determined, and of definitions and assumptions whose formalization is not completely correct with respect to the concept they refer to (see also the case of discriminating algorithms discussed below). Z. Lipton clearly explains that often machine learning problem formulations are imperfect matches for the real-life tasks they are meant to solve[4]. This can happen when complex real-life goals are difficult to encode as metrics or simple numerical functions to be optimized. For instance, ethics and legality cannot be directly reduced to numerical optimization objectives of a decision-making algorithm. When algorithms deal with goals that we deem important but

[2] Russell & Norvig 2010

[3] Russell & Norvig 2010

[4] Lipton 2016

struggle to model formally, problematic requirements like interpretability, explanation, transparency are called for[5]. We think that the knowledge developed by humanities can be useful here, since the research method of these disciplines has a clear notion of what it means to be *precise* without being mathematically *formal.*

2.2. A single language or many languages that interoperate?

An effective dialogue between different disciplines requires a shared understanding of the main notions of the topic under discussion. In the case of AI, we can list for example the terms *intelligence, behaviour, will, similarity, causality, neuron, action, accuracy, truth, fairness, precision*, but many other could be put forward. Each of these terms belongs to the vocabulary of multiple different disciplines, where it denotes different things and acquired specific connotations, nuances and possibly hidden references that became implicit in specific research communities. For instance, the terms *intelligence* or *causality* are used by computer scientists in a much narrower sense than by philosophers or neuroscientists. On the other hand, the terms *accuracy, precision* and *fairness* are used in machine learning algorithms with reference to very specific mathematical definitions, which valuably allows one to properly compare the performances of different algorithms[6]. Moreover, AI researchers sometimes use anthropomorphic terms and suggestive colloquial definitions (like *reading comprehension* algorithm or *thought vector*) that might be a fruitful source of inspiration when kept within the research community together with their proper technical qualification, but that can be confusing and give a misleading sense of the AI capabilities when communicated outside their original context[7].

Therefore, the difficult process of establishing a cross-disciplinary dialogue based on a common language has the advantage of bringing to light assumptions and connotations that might have become implicit, and possibly neglected or forgotten, in the day-to-day research jargon. However, for many concepts it might be impossible for different experts to completely agree on a common meaning without sacrificing the intended expressivity of a specific term. Therefore, instead of a single common language, a useful cross-disciplinary dialogue could be based on multiple discipline-specific languages that productively interoperate.

As an example, let consider the case of *disparate learning processes* (DLPs), which is a class of machine learning algorithms that has been put forward

[5] Lipton 2016; Mittelstadt, Floridi & Wachter 2017

[6] Russell & Norvig 2010

[7] Lipton & Steinhardt 2018

to address the discrimination issue of classification algorithms. Computer scientists resorted to a well-known legal terminology to define the technical criteria quantifying the algorithm's discrimination. More precisely, the legal notion of *disparate treatment* is a form of intentional difference in the treatment of protected subgroups, while a *disparate impact* refers to facially neutral practices that have unequal outcomes because of implicit correlations between protected an unprotected characteristics of individuals. Similarly, a classifier algorithm is said to avoid disparate treatment if it is blind to the protected characteristics of the input data, while its impact disparity is measured by checking whether the proportion assigned to the positive decision is equal across different groups of individuals. It turns out that DLPs algorithms satisfy both technical criteria, but a judge in a court would assign to DLPs algorithms no better legal status than explicit treatment disparity, since they essentially achieve group parity at the cost of individual unfairness[8]. Therefore, while the technical terms are inspired by legal concepts, the plain optimization of technical criteria may fail to satisfy the legal and ethical desiderata underlying the legal criteria. This example illustrates the difficulty of communicating desiderata across different disciplines, and shows a case where it is important to maintain the distinction between technical and legal terminology, finding instead a way to let the two languages fruitfully interoperate.

2.3. Identifying the levels of abstraction

An important source of weakness in the development of digital tools based on AI is that it is easy to blur different abstraction levels, thereby getting stuck with a problem that should have been addressed at a different development stage. The most prominent example is the distinction between the *classification* and the *decision-making* tasks in machine learning algorithms. Even if the automated output decisions are taken on the basis of the automated classification, these two tasks generally operate under different constraints and different goals, thus they should be designed, implemented and optimized by taking care of these differences. For instance, consider the case of DLPs algorithms discussed above, a fair solution could be obtained by letting the classification task perform an unconstrained learning approach even if the resulting outcome reflects a historical prejudice encoded in the input data. Such a classification model is unfair but has the advantage of being more transparent than that of DLPs, thus it leaves open the possibility

[8] Lipton, Chouldechova & McAuley 2018

of intervening at decision time to explicitly and transparently promote more equal outcomes aligning decisions with social desiderata[9].

Another important scenario where the distinction between classification and decision-making is crucial, is the autonomous car. The official preliminary report by the US National Transportation Safety Board on the fatal Uber self-driving car collision on March 2018[10] clearly shows that the autonomous classification system correctly identified the incoming obstacle and determined the need for an emergency braking maneuver. The report continues stating that "according to Uber, emergency braking maneuvers are not enabled while the vehicle is under computer control, to reduce the potential for erratic vehicle behaviour. The vehicle operator is relied on to intervene and take action. The system is not designed to alert the operator". A proper analysis of this accident would require a deeper discussion, however this paragraph from the official report shows that there is an issue in deciding what to do on the basis of the classification outcome. It also reveals that, since the Uber system is not designed to alert the operator, there is no alert even at the moment when the responsibility of decision-making is shifted from the machine to the human operator.

Generally speaking, the classification and prediction tasks are under the control of technical criteria, that mostly come from the theory of machine learning algorithms. Instead, the decision-making task entails much broader considerations, whose assessment depend (also) on social, ethical and legal criteria. By properly distinguishing these two phases, the difficulties of combining them more clearly emerge, and appropriate solutions may originate from the cross-disciplinary dialogue described above. Finally, medicine and public health is a domain where the application of AI and the distinction of the abstraction levels at which it operates are particularly delicate. In this context, the distinction between classification and decision-making turns into the distinction between, e.g., the classification of medical images and a diagnosis. Moreover, to make a diagnosis it is important to consider the difference between data and what data refer to; accordingly, a doctor knows how to interpret the numerical values of specific analyses in the light of the general anamnesis of the patient. Finally, two different abstraction levels pertain to data and to the methodologies of collecting, modeling, relating and examining these data. Many issues appear in each of these stages[11], bringing out the fundamental role of doctors' professional judgment, grounded on an ability to integrate facts and values, the demands

[9] Lipton, Chouldechova & McAuley 2018

[10] US National Transportation Safety Board 2018

[11] Cabitza, Rasoini & Gensini 2017; Cabitza 2017

of a particular case and prerogatives of society, and the delicate balance between mission and margin[12].

3. The dialogue between Science and Technology

Computer science, and artificial intelligence in particular, is a domain in which science and technology often intermingle, as testified also by the rich literature on philosophy of science devoted to AI. Without presuming to cover this topic here, we simply observe that the scientific research addresses knowledge while the technological development is devoted to building applications. These two paths are based on different methodologies, that can productively interoperate if, as in the case of the cross-disciplinary dialogue, they are combined so that cooperation and integration are obtained without losing their specificity.

A distinctive aspect of science is that its results are always open to be refuted, invalidated or subsumed by new results. The scientist investigates the limits of the knowledge, trying to find something new by questioning the established understanding and testing its robustness and its replicability. The technological development is rather devoted to take the most from an idea, a theory, a scientific result, with the aim of producing an artifact that is convenient according to some intended goal. It is worth observing that it's up to social discussion and politics to define a legal framework that marks the limits that technological artifacts should respect.

The relationship between science and technology is fundamental also to properly address the ethical questions raised by digital technologies. F. Russo calls for a rethinking of the relations between knowledge and its applications in order to avoid the so-called technological determinism, that is an either utopian or dystopian predefined path[13]. She proposes the *information ethics* framework, which is rooted in the philosophy of information and is based on the idea that we are not victims of technologies: we do not just build arguable digital artifacts, we also create the environments, possibilities or affordances, that are subject to ethical evaluation as well. We must therefore pay attention to which possibilities we decide to develop or not to develop, becoming responsible for the space of possibilities that we create.

Such a view brings to light different responsibilities involved at different abstraction levels, calling for a dialogue also between the ethics of the scientists and the ethics of engineers. As far as AI is concerned, new kinds of research papers and workshops are emerging to host such a dialogue, debating

[12] Pasquale 2019

[13] Russo 2018

arguments and points of view on major issues within the field and around the future of the AI technology; they can be viewed as concrete examples of ethics at work. For instance, G. Marcus provides for a critical reflection on the state of the art of deep learning systems, putting forward impressive advances, weaknesses and common misunderstandings[14]. N. Japkowicz and M. Shah point out that the performance evaluation of a machine learning algorithm is not just a matter of applying the correct mathematical formula, but is also a problem of appropriateness of the chosen evaluation method and interpretation of the results obtained[15]. Furthermore, Z. Lipton and J. Steinhardt review the scientific literature on machine learning putting forward a number of troubling trends that hinder the future research and compromise AI's intellectual foundations[16]. The flawed patterns singled out from research papers are the failure to distinguish between explanation and speculation, the failure to identify the correct sources of empirical gains, the use of mathematics in a way that obfuscates or impresses rather than clarifying, and the misuse of language by choosing terms with colloquial connotations or by overloading established technical terms. Taking a constructive attitude, the authors also speculate on the possible causes behind the problematic trends and provide a discussion about what the research community can do to raise the level of experimental practice, exposition, and theory, and to disabuse researchers and the wider public of misconceptions[17].

4. The dialogue between Research and Society

Finally, the complexity of the impact of AI systems on society and on people's lives asks for an earnest dialogue between research and society. Most modern digital systems are better qualified as *socio-technical* systems, since their technical design is affected by and has an impact on the behaviour of users. These systems actually provide for infrastructural services in the domains of production, business, communication, entertainment, education, city planning, access to health, up to access to democracy and human rights exercise. As put forward by K. Crawford[18], today AI is three things: (i) a set of technical approaches, (ii) a set of social practices that powerfully shape the AI systems according to non-technical decisions, like who works on these systems, who decides which problem is prioritized, how humans would

[14] Marcus 2018

[15] Japkowicz & Shah 2011

[16] Lipton & Steinhardt 2018

[17] Lipton & Steinhardt 2018

[18] Crawford 2018

be classified, and (iii) a profoundly concentrated industrial infrastructure. Therefore researchers have a duty to recognize the social, political and ethical natures of technological artifacts, and they should engage with the audience for their research papers, which has broadened so to increasingly include students, journalists, and policy-makers.

To conclude, we need both a more socially literate scientific community and a more scientifically literate public[19]. To this end, scientists should be effective communicators of scientific issues, making understandable the working principles of digital systems and their consequences on users. This is especially important in machine learning systems that include classification and decision-making algorithms that can hardly be interpreted by non experts; see for instance the European recommendations on machine-learned automated decision making provided by the Informatics Europe and EU-ACM scientific associations[20]. On the other hand, society needs to dialogue with the scientific community, first of all to challenge the degraded state of public discourse on science. Then a social conversation about shared values and shared objectives is necessary to effectively press and orient the priorities of technical development.

5. Conclusions

The substantial advances in the performances of Artificial Intelligence applications gave rise to an increased enthusiasm and an exceptional level of attention outside the scientific community. Both the opportunities and the concerns that AI brings about require the involvement of experts coming from many different domains: scientific, technological, social, ethical, legal, economical, psychological, political, educational, artistic. However, a multidisciplinary approach requires an effective communication between people that have different backgrounds, use different working methodologies and essentially speak different languages. When different machines communicate, they are said to *interact*, while a communication between different people is a *dialogue*. Such a difference brings in the richness of human intelligence, which does not simply process information, but operates on ideas, intuitions and even emotions. Therefore, a dialogue between people with different expertise and different roles in the society makes it possible to move from a multidisciplinary approach toward a cross-disciplinary, deeper, investigation. In this article we discussed many facets of

[19] Hoffman 2016

[20] Informatics Europe & EU-ACM 2018

these dialogues, showing how different approaches and methodologies can cooperate and produce new insights without losing their specificity.

It remains problematic to understand how to develop such an effective cross-disciplinary dialogue[21]. We think that in this challenge the psychological and pedagogical ideas of John Dewey[22] can be enlightening: with the so-called *learning by doing* approach, the American philosopher emphasized the role of active experiences in grasping the meaning of concepts. In his view, learning is always an interactive and social process, because the concrete transmission of knowledge takes place through shared experiences, where words more explicitly show the meaning for which they are used, which we have seen is a particularly subtle aspect when it comes to cross-disciplinarity. Moreover, according to Dewey, the learning method should not be imposed nor be hierarchical (thus assigning to a discipline a priority over another one), but cooperative and democratic, in analogy to the spirit of the scientific method, which is made of verification, criticism and sharing, aimed at the growth of the body of knowledge. We think that the learning by doing methodology and Dewey's ideas on the role of learning and education on the social progress can be very effective in the case of AI, as testified by the lively and engaging cross-disciplinary discussions that inspired this article.

Acknowledgements

I am grateful to S. Fuselli, P. Moro, C. Sarra, A. Testolin, M. Zorzi and L. Pelizzon for valuable and insightful cross-disciplinary discussions, that gave rise to the arguments systematized in Section 2 and inspired the article's perspective.

References

US National Transportation Safety Board 2018. *Preliminary Report High-way HWY18MH010*. Arizona, USA, 24th May 2018. https://www.ntsb.gov/investigations/AccidentReports/Pages/HWY18MH010- prelim.aspx.

Cabitza, F. 2017. *Breeding electric zebras in the fields of medicine*. CoRR, abs/1701.04077 and IEEE workshop on the Human Use of Machine Learning (HUML 2016).

[21] Crafa & Pelizzon 2018

[22] Dewey 1916

Cabitza, F., Rasoini, R. & Gensini, G. 2017. “Unintended consequences of machine learning in medicine”. *The Journal of the American Medical Association (JAMA)*, 318(6):517–518.

Crafa, S. & Pelizzon, L. 2018. *Epistemological questions for a philosophical education in artificial intelligence.* SILF (Societa Italiana di Logica e Filosofia della Scienza) Post-graduate Conference. 2019.

Crawford, K. 2018. *You and AI - just an engineer: the politics of AI.* Distinguished lecture, Royal Society, London, https://www.youtube.com/watch?v=HPopJb5aDyA.

Dewey, J. 1916. *Democracy and Education.* New York: Macmillan.

European Parliament 2017. Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

Harwell, D. 2018. *Unproven facial-recognition companies target schools, promising an end to shootings.* The Washington Post.

Hoffman, A.J. 2016. “*R*eflections: academia’s emerging crisis of relevance and the consequent role of the engaged scholar”. *Journal of Change Management* 16(2):77–96.

Informatics Europe & EUACM. 2018. *When computers decide: European recommendations on machine-learned automated decision making.* http://www.informatics- europe.org/working-groups/ethics.html.

Japkowicz, N. & Shah, M. 2011. *Evaluating Learning Algorithms: A Classification Perspective.* Cambridge: Cambridge University Press.

Lipton, Z. 2016. *The mythos of model interpretability.* ICML Workshop on Human Interpretability, and Communications of the ACM 61(10):36-43, 2018.

Lipton, L., Chouldechova, A. & McAuley, J. 2018. *Does mitigating ML’s impact disparity require treatment disparity?* Conference on Neural Information Processing Systems (NeurIPS).

Lipton, Z. & Steinhardt, J. 2018. *Troubling trends in machine learning scholarship.* CoRR, abs/1807.03341.

Marcus, G. 2018. *Deep learning: A critical appraisal.* CoRR, abs/1801.00631.

Mittelstadt, B., Floridi, L. & Wachter, S. 2017. “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7(2):76–99.

Pasquale, F. 2019. “Professional Judgment in an Era of Artificial Intelligence and Machine Learning”. Boundary 2, 46(1):73–101.

Powles, J. 2017. *New York City's bold, flawed attempt to make algorithms accountable.* The New Yorker.

Russell, S. & Norvig, P. 2010. *Artificial Intelligence: a modern approach.* New Jersey: Pearson.

Russo, F. 2018. "Digital technologies, ethical questions, and the need of an informational framework". Philosophy & Technology 31:655-677.

United Nations Human Rights Office of the High Commissioner. 2018. *Statement on visit to the United Kingdom, by professor Philip Alston, United Nations Special Rapporteur on extreme poverty and human rights.* https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23881&LangID=E