Explanation for Humans, for Machines, for Human-Machine Interactions?

Rémy Chaput, Amélie Cordier, Alain Mille

Explanation for Humans, for Machines, for Human-Machine Interactions?

Rémy Chaput,[1] **Amélie Cordier,**[1,3] **Alain Mille**[1,2]
[1] Université de Lyon, Université Lyon1, LIRIS UMR CNRS 5205
[2] Coexiscience, Lyon, France
[3] Lyon-iS-Ai, Lyon, France
remy.chaput@univ-lyon1.fr, amelie@lyonisai.fr, alain.mille@univ-lyon1.fr, alain.mille@coexiscience.fr

Abstract

The XAI concept was launched by the DARPA in 2016 in the context of model learning from data with deep learning methods. Although the machine learning community quickly took up on the topic, other communities have also included explanation in their research agenda (e.g. Case Based Reasoning, Planning, Decision Support, Emerging Systems, Robotics, Internet of Things). The question of explanation, which is at the center of philosophical research works, has been revisited during the last decades. The humanities community insists on the fact that explanation is above all a process that develops in the context of the search for explanation and cannot be completely defined *a priori*. In this contribution, we propose 1) to broaden the question of explanation to any type of situation in which users exploit the possibilities of decision support agents for their own decisions, in the context of their task, and within the framework of their activities and responsibilities and 2) to consider an instrumentation of digital devices, able to manage dynamic explanation agents associated to corresponding decision support agents. We denote this evolution "UXAI" (User eXplainable Artificial Intelligence) because we consider that users should be the main actors in the dynamics of any explanation process.

Introduction

The international research movement on "Explainable Artificial Intelligence" (XAI) has grown considerably over the last years. Literature on the subject is abundant. See, for example, (Mueller et al. 2019; Anjomshoae et al. 2019). The urgency of this research is revealed with the extraordinary impact in society of "artificial intelligence" and even more so when these "AI" are the result of learning from massive data, and even more so, if possible, when the learning exploits so-called "deep learning" techniques. This demand for explanation by society can be explained by:

- Ethical reasons (Bird et al. 2020),
- Reasons related to the establishment of responsibilities: the question arises, for example, with the use of COMPAS in the USA,
- Reasons related to the semantics of models produced by data analysis to discover laws of the world (van der Schaar 2020),
- Economic reasons as outlined in the MIT report (Gunning and Aha 2019).

The kick-off for a specific research field has been made by the Defense Advanced Research Projects Agency (DARPA), when publishing, in 2016, an announcement focusing on numerical machine learning (Gunning and Aha 2019). This report explains the issues with a situational scenario which unfolds as follows (sic.).

- **Today**: the users of a function learned from data asks themselves questions to which there are no possible answers: Why did you do this? Why not something else? When do you succeed? When do you fail? When can I trust you? How do I correct an error?
- **Tomorrow**: the user with an explanation interface: I understand why. I understand why not. I know when you succeed. I know when you fail. I know when to trust you. I know why you erred.

The XAI machine learning community completed the specifications by listing the involved audiences (Barredo Arrieta et al. 2020): experts in the field so that they can evaluate the model and acquire scientific knowledge; regulatory institutions and agencies so that they can certify compliance with legislation, regulations, audits, etc.; managers and board members to assess the model's compliance with their regulations and understand their company's AI applications, etc.; scientific data specialists, developers, product owners so they can verify and improve the effectiveness of products, research, add new features, etc.; users affected by decision models to understand their situation, verify the fairness of decisions, etc.

In this article we posit that these questions arise regardless of the technologies used for the development of digital devices intended to "help the decision".

Classical technologies, with the notable exception of web technologies, encapsulate regulations in closed boxes, without any real possibility of sharing the associated knowledge with the end user. We posit that artificial intelligence technologies should rather be a chance to facilitate explanatory processes by designing agents in this way. The case of model learning by deep learning raises the question in a new way since even the designers developing the algorithms generating the learned models do not have access to the semantics of the decisions that these models will support.

We also raise the fact that if numerous research works take into account the end users, the goal is most often to establish and integrate their profile to adapt pre-constructed explanations. The fact that the user can be the main actor and co-constructor of the explanation is seldom thought of. Only a few research works in the field of the humanities and social sciences really address this question (Miller 2019). Humans introduce explanatory biases (de Graaf and Malle 2017) as they project their own explanatory patterns on artificial agents, whether materialized or not. The explanation process must reduce these biases by providing formal and non-emotional explanatory patterns.

Finally, we posit that an explanation is a complex process that is co-constructed with the users in the context of their tasks, their responsibilities, their knowledge, their abilities to deliberate decisions alone or with others. We propose to explore a new research perspective, UXAI (User Explained Artificial Intelligence), grounded on these claims.

In this article, we questions different artificial intelligence technologies on their capacities to facilitate an explanation process for end users in a contextualized decision situation. This article ends by opening avenues to answer the question of how to manage the process of explanation, by facilitating the interaction between humans and digital device, when these devices are used for decision support and must make it possible to reveal the actual working regulations.

Explanation: what does it mean for humans?

Philosophy questions the notion of explanation through a *theory of explanation*. This theory first focused on how to establish causes to phenomena and observed things. How to answer to the question WHY? The causal approach remains a lively one when establishing scientific laws from massive collected data. At the beginning of the 20th century, major scientific progresses, particularly in physics, were made without calling into question the principle of establishing causality based on direct observation. Physicists would formalize laws that explain macroscopic or microscopic effects impossible to observe in isolation, and also explains effects whose causality could be established by observation. So formalized laws describe a *reality* that goes beyond what is observable by the usual methods. The study of the *theory of explanation* is then declined according to a realistic (empirical) or epistemic orientation (Galavotti 2018). For a realistic (empirical) approach, an explanation is a literal description of an external reality. For an epistemic approach, the explanation is used to facilitate the construction of a coherent empirical model as stated by Bas van Fraassen[1]. Explanation theory also integrates the study of the process of explanation. For example, the philosophy of language focuses on understanding between individuals (Achinstein 1985), while cognitive sciences assert that explanation is above all cognitive and results from a mental representation linked to the activity and helping this activity (Mayes 2020; Holland et al. 1987; Horne, Muradoglu, and Cimpian 2019). The interested reader can find detailed elements in the long-term study conducted by Robert R-Hoffman in a series called Explaining Explanation (Hoffman and Klein 2017; Hoffman, Mueller, and Klein 2017; Hoffman et al. 2018) which ends with a specific study of machine learning by G. Klein (Klein 2018). In Explaining Explanation for "Explainable AI" (Robert R. Hoffman, Klein, and Mueller 2018), the authors point out the elements that make "a good explanation" within the framework of devices such as Artificial Intelligence. These elements constitute a theoretical contextualization to the series of articles presented above. According to Hoffman, the key observations to study the notion of explanation for humans are the following. (1) Explaining is a continuous process: Humans are motivated to "understand the goals, the intention, the awareness of the context, the limitations of the task, [and] the basis for analyzing the system to see if it can be trusted" (Lyons et al. 2017); (2) Explaining is a co-adaptive process: "Explanations improve cooperation, cooperation allows the production of relevant explanations" (Brezillon and Pomerol 1997); (3) Explanation must be triggered: think about what triggers the need for an explanation, which can be based on the phenomenon of "surprise" in the face of an unfulfilled expectation, for example; (4) We should facilitate self-explanation: self-explanation is the fact of finding the explanation with or without help; (5) Explanation is an exploration: the account of the exploration is part of the explanation, showing the path of questions, answers, and the completion of information to answer the questions (Mueller and Klein 2011); (6) Contrast situations can be explained differently (Miller, Howe, and Sonenberg 2017).

[1]For details, see: https://fr.wikipedia.org/wiki/Bas_van_Fraassen

Explanation for AI based machines?

In this section, beyond the specific explanatory capabilities of an AI agent, we address the requirements of an explanation process that includes the user in its unfolding. We point out the strengths and weaknesses of different types of AI agents when contributing to such a process.

Explanation and symbolic based expert systems

Expert systems represent knowledge, rules and facts in a symbolic form. Reasoning uses rules to infer new facts from established facts. Reasoning is traceable, and would therefore, by its nature, be explainable. However, after the euphoria, expert systems have left the limelight. (Swartout and Moore 1985) analyzed the reasons for this failure. The explainability of these systems was judged to be weak and constituted a significant reason for their abandonment. Mycin (Shortlife et al. 1975), a famous medical diagnostic system, demonstrated superior decision-making qualities to physicians and even to groups of physicians. The inability of physician users to integrate their own knowledge and contextual knowledge into the system distanced them from the benefit of diagnostic quality, which, indeed, they did not dispute. Case-Based Reasoning (CBR), in which the learning loop integrates users, organizes knowledge into cases, a representation similar to users' practices. CBR is more effective and has played an important role in the management of practical knowledge in companies. This no doubt explains why

the Case-Based Reasoning community is particularly active in the XAI domain. Since Alan Newell's proposal (Newell 1982) to separate the representation of knowledge from its exploitation mechanisms, an important research community has developed around ontologies and knowledge engineering in general (and more recently, around the semantic web). With ontologies, many hoped that we had found the absolute tool to explain the knowledge mobilized in a decision. However, the construction of these knowledge graphs is difficult and normalizing. Knowledge graphs are often created from available corpora, especially on the web (Biemann 2005). Many decision-support devices are still based on explicit and symbolic knowledge, with or without probabilistic, fuzzy or possibilistic moderators. Different modalities of logics have been experimented to improve the adaptation to real situations, without any better explanatory success. To our knowledge, no research has been published that considers the explanation of expert systems as a process that requires giving the user the central role for learning to explain by integrating the contextual elements of the decision to be supported.

(Chakraborti, Sreedharan, and Kambhampati 2020) is one of the first papers to express the concept of *explanation process* within the community of Planning and Decision Helping. They use the term of *emerging landscape* for *having to explain its decision can be folded into an agent's reasoning stage itself*. A number of recent papers demonstrate a very high interest for *XAI* in the classical scope of planning and decision helping (Magazzeni et al. 2018; Chakraborti et al. 2019).

Explanation and machine learning based systems

Machine Learning (ML) based systems intend to learn a numeric model from a large quantity of data; the model is later used to predict new data or to act in an environment. Recent works in ML, and particularly in Deep Learning (DL) have shown remarkable results, sometimes exceeding human performance (for example on image recognition tasks (Zhang et al. 2018)). However, ML models can also make trivial mistakes, such as labeling an axe as a screwdriver (Hoffman et al. 2018).

As such, an explanation process may serve two main objectives: first, to allow end users who use learned models to understand the "reasoning" which led to a decision so as to accept the decision or not; and secondly, to allow designers or regulators to understand the underlying causes of a mistake so as to refine and correct the model.

The wide diversity of existing ML methods calls for a variety of mechanisms to produce explanations, which can be placed along three axes. First, the mechanism can produce *post-hoc* explanation after the learning occurred, or an *intrinsic* explainable capability of the learning algorithm. Secondly, the explanations can be either *global*, i.e. focusing all instances of the dataset, or *local*, focusing only a specific instance. And finally, they can rely on the model's internal characteristics (*model-specific*) or not (*model-agnostic*).

We detail below a categorization of 4 commonly used types of problems (Guidotti et al. 2018).

Model Explanation Considering a non-easily understandable model, named the "black-box", and a set of test instances for which we do not have access to a ground truth, the Model Explanation problem aims at learning a second, "transparent" model for the testing data using the inputs and outputs of the black-box model. This type of method provides *post-hoc*, *global* explanations and, more often than not, in a *model-agnostic* fashion. As the resulting "transparent" model must be easily understandable by humans, several authors recommend using Decision Trees; however, one can wonder to which extent this model provides a satisfying explanation process. Indeed, when the "transparent" model is model-agnostic, the explanation may not be coherent with the actual "reasoning" of the black-box model (i.e., there is no fidelity). Moreover, this assumes that the audience will be able to understand such "transparent" models; this depends on the size of the generated Decision Tree for example, and the expertise of the audience. Finally, this method only provides an explanation, and not a complete explanation process; in other terms, the user cannot interact with the explanation and ask for more details.

Outcome Explanation Contrary to the first problem, the Outcome Explanation problem focuses on providing an explanation for a single instance, also called "local explanation" (Barredo Arrieta et al. 2020). This problem may use a similar method to the first problem, by generating a "transparent" predictor, specific to this instance; it is also possible to generate counterfactuals, i.e., foil data that differs slightly with the real input data but for which the output is different to allow the user to compare. However, such explanations suffer from the same criticisms as the first problem: the fidelity to the black-box's actual algorithm is not guaranteed, and this does not constitute an explanation process.

Model Inspection The Model Inspection problem focuses on the internal characteristics of the model and of the dataset. As such, it pertains to *intrinsic* explanations instead of *post-hoc* as previously; both *global* and *local* explanations can be generated. Methods that purely rely on the dataset, e.g. Principal Component Analysis, are *model-agnostic* but other methods may explain characteristics of the model and therefore are classified as *model-specific*. While this problem allows understanding which variables affect the model's prediction and to which extent, one cannot understand the reasons of this influence.

Transparent Box Design The Transparent Box Design problem is fundamentally different from the first three, as it focuses on intrinsic explanations. Instead of providing explanations for a black-box model, the goal is to directly produce a "transparent" model from the dataset. A part of the XAI community is increasingly putting this type of methods forth, in particular for applications with important stakes (Rudin 2019). Although "transparent" models have advantages, such as allowing one to easily detect bias in the produced rules, the result is not a satisfying explanation process, as there is no real explanation and no construction of the explanation with a given user.

Toolkits Following the upsurge of Explainable AI articles focusing on Deep Learning models, industries recently proposed toolkits to help developers easily integrate elements of explanation into their applications. However, these solutions are (for the time being) still limited in terms of adaptation to the audience and interaction.

We briefly describe two of the proposed toolkits: TensorFlow What-If Tool[2] (WIT) and IBM AI Explainability 360[3] (AIX 360).

The TensorFlow WIT allows to explore the dataset and to visualize the importance of variables. It is also possible to temporarily change a variable from a datapoint to create a counterfactual; the user is therefore able to interact with the model to create its own mental representation. This interaction capability is the main advantage of WIT, however, the tool lacks adaptation to the audience. Indeed, the datapoints are displayed as raw variables and are therefore targeted towards AI developers, data scientists or potentially domain experts; it seems unlikely that end users may benefit from such an interface.

On the other hand, IBM AIX360 offers a multitude of methods to produce explanations, each of them targeting a different goal. For example, the toolkit proposes methods for local or global explanations, using an interpretable model or post-hoc explanations. Multiple types of audience can be targeted; as such, it is possible to adapt the explanation by implementing several methods. However, the demonstration show little to no interaction: for example, a bank client may see which variables are important to get its loan file accepted, but cannot simulate foil data to observe the effect of changing a specific variable.

Explanation and robotics/IoT based systems

Although AI is a discipline that is almost 70 years old, it looks like the concerns about explanations are only very recent. If we take a step back, and look at the phenomenon not only from a scientific point of view, but also from a societal point of view, this observation justifies itself quite easily. Of course, we can consider the willingness to make explainable AIs as a kind of response to the myth of the black box of the machine learning, but this is not the only reason (and Machine Learning does not imply black boxes, it implies models whose explanation is possible, but not accessible to humans). The need for explanation is way more important than that: understandable explanations are the absolute condition for the acceptance of artificial intelligence. Without actionable explanations, humans won't trust AIs, and more importantly, won't be able to use them wisely.

This need for explanations grows fast because, it is not anymore to prove, AI has become incredibly present in almost all business areas. Companies, consultants, analysts, journalists, and even institutions (like the European Commission) are all advocating for better explanations, each of them providing different, but converging, arguments.

However, most of the explanations that we are currently able to provide do not really resemble the explanations that a human would give to justify their choices. When humans explain things to each other, they have a natural tendency to align themselves with each other, according to their knowledge, the vocabulary they are used to using, etc. They rarely (if never) describe, step by step, the different phases of reasoning they followed in reaching their decisions or conclusions.

In the field of robotics (de Graaf and Malle 2017), and more precisely in social robotics and of collaborative robotics, the need for explanation is usually a main concern. Indeed, robots are designed to interact with human beings in a professional context, and everything must therefore be implemented to ensure that humans trust robots they are working with, and to make sure that the robots can operate safely.

As an example, we can cite several European projects that stand out in these areas and who have established working groups or expert committees to deepen the topics related to explanations. The Crowdbot project[4] for example, whose objective is to allow robots to freely navigate among humans, places the issues of security and robustness at the heart of their design.

To give another example in robotics, the Inbots project[5], which aims at providing inclusive robots for a better society, has set up a number of committees of experts to study the societal and socio-economic uptake of the work they carry out.

In the field of Internet of Things (IoT), the explainability of systems is also a frequently observed issue. However, it takes a different dimension. Whereas many people express concerns about safety, security, data privacy and resilience of IoT devices, fewer are those who understand that the difficulty in explaining IoT systems comes from the complexity that emerges from a vast network of simple devices.

Explanation is not only a matter of providing details on how a single device works, but it also means being able to explain the interactions between the devices, and the chain of decisions.

For example, the paper (Garcia-Magarino, Muttukrishnan, and Lloret 2019) describes a use case inspired from the everyday life (a connected kitchen) to demonstrate how IoT and AI can be combined to provide explanations. In a completely different field, the paper (Tsakiridis et al. 2020) describes an explainable approach to AI in the field of farming.

Explanation and bio-inspired systems

Bio-inspired and emergent systems (Bonabeau, Dessalles, and Grumbach 1995; Darwish 2018) are a class of mechanisms studied to solve optimization problems by simulating behaviors models observed in nature. Building numerical models from observed behaviors is a difficult task even though a number of meta-models have been described for types of behaviors (Ferber and Gutknecht 1998). Many of these systems are dedicated to optimization research or

[2]https://www.tensorflow.org/tensorboard/what_if_tool

[3]https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

[4]http://crowdbot.eu/

[5]http://inbots.eu/

to the discovery of potential behaviors according to specific meta-heuristics. These models are sometimes considered simplistic and more radical approaches are proposed to better account for natural emergence models (Di Paolo and Lizuka 2008). For the past two years, the international reference conference AAMAS has been hosting an Extraamas workshop around XAI (Calvaresi et al. 2019, 2020). The objectives of the workshop are indeed broad: "(i) to strengthen the common ground for the study and development of explainable and understandable autonomous agents, robots and Multi-Agent Systems (MAS), (ii) to investigate the potential of agent-based systems in the development of personalized user-aware explainable AI, (iii) to assess the impact of transparent and explained solutions on the user/agent behaviors, (iv) to discuss motivating examples and concrete applications in which the lack of explainability leads to problems, which would be resolved by explainability, and (v) to assess and discuss the first demonstrators and proof of concepts paving the way for the next generation systems." The contributions mainly concern (ii) and (iii) with some exceptions on (i) (Alzetta et al. 2020) but the question of the transparency of a device for the end user is also addressed in its complexity (Tulli et al. 2019). From a certain point of view, it is indeed a question of using an artificial intelligence technique to facilitate an explanation process. The focus of the research seems targeted at the process of elaborating an explanation in a complex situation, with multiple models and associating both artificial and human agents.

Towards an UXAI model for dynamic explanation processes?

The detailed study of the state of the art shows the weakness of the consideration of explanatory processes in XAI, and that these explanatory processes are important to consider in the different fields of Artificial Intelligence, not only for deep learning based approaches. This leads us to affirm that: 1) even if deep learning raises specific problems, the question of explanation arises for any numerical form of decision support agent. Actually, the fact that digital agents rely on artificial intelligence techniques should facilitate the implementation of a dynamic explanation process whose main actors are user in their contexts of use. 2) It is possible to design a dynamic explanation process based on explanatory agents that are sufficiently "intelligent" to learn with users how to explain the behavior of a numerical decision support agent. We propose a scenario for the use of such an approach that we call UXAI (User Explained Artificial Intelligence), and compare today's and tomorrow's situations from this perspective. We make some hypotheses on how to achieve this.

Today

Research Level Researchers build decision models. They do this from their expert knowledge and from collected data. The collected data is used to check the validity of the models but also, and dramatically, to learn the models. It is within the framework of the automatic learning of models from data that the research called XAI was launched and is giving its first results. We call General Model (GM) a model as published at this level. These are algorithms for building applied models or generic application models trained specifically for families of applications (for example: Convolutional Neural Networks for Computer Vision).

Production Level Application design and deployment use published general models to build applied models, which are operationalized in the context of an activity with specific objectives for the deployment of applications on real-world digital devices (as Face Recognition). Producers (developers, designers, marketing people, etc.) mobilize specific knowledge for their objectives and specialize the general models to their context of use. The user experience they seek to satisfy or they would like to propose is part of the mobilized knowledge. The information that feeds their models comes from ad hoc data collection (Mechanical Turk for example) or from the work of the designers (use cases). We call Applied Model (AM) a model designed at this level. The AM is implemented in the form of encapsulated functions in the applications as specified by the designers.

Usage level Users are using the application as a support for their activity under their responsibility. Users mobilize their own knowledge to use the application in order to carry out their task in a real context. Interaction traces provide data that can be used to specialize the functioning of the applied model, but also, and increasingly, to provide information to producers to evolve the applied model itself. There is already a loop with the producer level. Researchers can also collect data from the implementation of applied models to build the data corpuses that feed into their own model design or model learning activity.

Tomorrow, with UXAI

Research level Researchers associate general models of explanation (UXAI-GM) with their general models (GM), gradually incorporating the necessary knowledge to be disclosed as an explanation so that the explanation process with users can take place in a more informed manner. To achieve this, they can work with feedback from designers in the form of specifications of explanations to be produced together with the models they build and, in a more upstream approach, they can anticipate and include potential users in action research sequences on the field of usage. They derive from this research the necessary concepts to propose general user oriented explanation. They focus specifically on producing general explanation-oriented models with the user (UXAI-GM). One research challenge is to join the production of a GM model with its corresponding UXAI-GM. Research methods and protocols are impacted, including in the editorial charts of publications, such as ethical requirements for publishing research works.

Production level Designers consider users not only when producing the explanation models (UX approach) that they deploy, but integrate in these models the possibility of co-constructing the explanations with the users themselves. A first UXAI-AM is designed from a UXAI-GM considered by the designers. It will be installed at the same time as the AM.

However, this model has to be customizable by the users in their own contexts. More precisely, UXAI-AM type models could be designed as "intelligent" agents with associated knowledge and explanatory inference mechanisms. For example, studies show (Wortham, Theodorou, and Bryson 2016), including to understand the behavior of robots (Potochnik 2011), that it was necessary to provide users seeking to understand what is going on, a form of trace of the internal states of the artificial intelligence they use. This UXAI dynamic is directly related to the ability to conduct research to study the architecture and knowledge learning mechanisms for these models. This research closely involves designers. One research challenge is to make the evolution of an AM and the corresponding UXAI-AM synchronous. Application production chains are concerned by this requirement. The validation of an AM is linked to the validation of the corresponding UXAI-AM.

Usage level It is at the user level that change is most evident. The user has not only an application but also an explanation agent associated with it: the UXAI-AM model and its explanation knowledge. The user can then learn to understand the behavior of the application with the help of the associated explanation agent. This explanation agent capitalizes the explanation processes in the form of an explanation learning memory. This memory associates the contributions of the users explaining in their own way and their own contexts with more generic explanatory knowledge provided with the explanation models. A situated explanation of the application is then possible and the explanation agent can follow this new way for reusing in comparable situations or for sharing with other users. If desired, users can easily share with designers to evolve their models. In turn, designers can share their knowledge with researchers to evolve existing general models of explanation. At this level, UXAI research mobilizes a wide variety of disciplines, in particular cognitive sciences, cognitive psychology, human-computer interactions, etc.

A simple example **Today**, let us imagine an application for risk assessment of bank loans. Researchers have proposed various decision models (decision trees, Deep Learning methods, etc.) which can be implemented by the bank. The bank's developers also add explanations for two different, specific audiences: their employees (who are domain experts), and the lay user; explanations tailored for these audiences have been also proposed by researchers, let us suppose for example that developers implemented a certain sort of variable explanation. When giving a user profile to the model, the result could be, e.g., "Rejected - The salary is not sufficient", or even "If the salary was 1500 instead of 1200, the result would be Accepted". This is an explanation, specialized for a given audience, however, it is possible that the users are not satisfied by this explanation. Perhaps they would like to know why does the model use 1500 as a threshold? The current explanations, although thought with the user in mind and made for users, are not made with the users. **Tomorrow**, let us imagine the same application, using our UXAI model: the UXAI-AM still provides explanations to the users, but also accepts explanations from users in some sort of interactive dialog. E.g., users could rephrase according to their own comprehension "So, I am too poor to be given a loan?", and the UXAI-AM could answer "No, but there is much more negative examples, where loans were not fully reimbursed, than positive examples. Therefore, the risk was deemed too high for the specified regulations." By allowing users to self-explain, and the UXAI-AM to correct these explanations, we can help users having a clear mental model of the decision.

Conclusion

The main reasons for the need for a process of explanation of digital decision support agents are mainly related to: the ethics of their use, the question of the respective responsibilities of the digital agent and its direct user; the economic question that could emerge if, unable to explain the decision, the user would prefer not to use these agents and even could consider it would be prudent to strictly regulate their use. The statement of these issues demonstrates that the user is at the center of the decision. It is the users' own ability to understand the workings of decision support that is required to enable their to evaluate and explain their own decisions in the context of their own context and responsibilities. After having explored the state of the art by opening it up to all types of artificial intelligence and having questioned more conceptually the notion of explanation as a process rather than as an element of information, we conclude by proposing to extend the research to the process of explanation and to unify it around a principle involving the user as the main and indisputable actor. We have named this orientation UXAI (User Explained Artificial Intelligence) to put the user at the start of any explanation process and to study what types of explanation assistance agents can then be developed to propose a concrete response to the requirement of mastering the human user's freedom of decision.

References

Achinstein, P. 1985. *The nature of explanation*. Oxford University Press. ISBN 0-19-503215-2.

Alzetta, F.; Giorgini, P.; Najjar, A.; Schumacher, M. I.; and Calvaresi, D. 2020. In-Time Explainability in Multi-Agent Systems: Challenges, Opportunities, and Roadmap. In Calvaresi, D.; Najjar, A.; Winikoff, M.; and Främling, K., eds., *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, volume 12175, 39–53. Cham: Springer International Publishing. ISBN 978-3-030-51923-0 978-3-030-51924-7. doi:10.1007/978-3-030-51924-7_3. URL http://link.springer.com/10.1007/978-3-030-51924-7_3. Series Title: Lecture Notes in Computer Science.

Anjomshoae, S.; Najjar, A.; Calvaresi, D.; and Främling, K. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceeding of AAMAS 2019*, 13–17. Montreal QC, Canada.

Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera,

F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115. ISSN 1566-2535. doi:10.1016/j.inffus.2019.12.012. URL http://www.sciencedirect.com/science/article/pii/S1566253519308103.

Biemann, C. 2005. Ontology Learning from Text: A Survey of Methods. *LDV-Forum* 20(2): 75–93.

Bird, E.; Fox-Skelly, J.; Jenner, N.; Larbey, R.; Weitkamp, E.; and Winfield, A. 2020. The ethics of artificial intelligence: Issues and initiatives. Scientific Foresight Unit (STOA) PE 634.452, Service de Recherche Parlementaire Européen (EPRS). URL https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf.

Bonabeau, ; Dessalles, J.-L.; and Grumbach, S. 1995. CHARACTERIZING EMERGENT PHENOMENA (2): A CONCEPTUAL FRAMEWORk. *Revue Internationale de Systèmique* 9(3): 347–371.

Brezillon, P.; and Pomerol, J.-C. 1997. Joint cognitive systems, cooperative systems and decision support systems: a cooperation in contexte. In *Proceedings of the European Conference on Cognitive Science*, 129–139. Manchester.

Calvaresi, D.; Najjar, A.; Främling, K.; and Winikoff, M. 2020. EXTRAAMAS2020. URL https://extraamas.ehealth.hevs.ch/.

Calvaresi, D.; Najjar, A.; Schumacher, M.; and Främling, K. 2019. *First International Workshop, EXTRAAMAS 2019, Explainable, Transparent, Autonomous Agents and Multi-Agent Systems*. Montreal QC, Canada, springer edition.

Chakraborti, T.; Dannenhauer, D.; Hoffman, J.; and Magazzeni, D., eds. 2019. *Proceedings of XAIP 2019 Explainable AI Planning*.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2020. The Emerging Landscape of Explainable AI Planning and Decision Making. *arXiv:2002.11697 [cs]* URL http://arxiv.org/abs/2002.11697. ArXiv: 2002.11697.

Darwish, A. 2018. Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications. *Future Computing and Informatics Journal* 3(2): 231–246. ISSN 23147288. doi:10.1016/j.fcij.2018.06.001. URL https://linkinghub.elsevier.com/retrieve/pii/S2314728818300631.

de Graaf, M. M.; and Malle, B. F. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). AAAI Technical Report FS n17-11, American Association for Artificial Intelligence. URL http://research.clps.brown.edu/SocCogSci/Publications/Pubs/de%20Graaf%20&%20Malle%20(2017)%20How%20people%20(AIS)%20explain%20action%20AAAI.pdf.

Di Paolo, E. A.; and Lizuka, H. 2008. How (not) to model autonomous behaviour. *Biosystems* 91(2): 409–423.

Ferber, J.; and Gutknecht, O. 1998. A meta-model for the analysis and design of organizations in multi-agent systems. In *Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160)*, 128–135. Paris, France: IEEE Comput. Soc. ISBN 978-0-8186-8500-2. doi:10.1109/ICMAS.1998.699041. URL https://ieeexplore.ieee.org/document/699041.

Galavotti, M. C. 2018. Wesley Samson. URL https://plato.stanford.edu/entries/wesley-salmon/.

Garcia-Magarino, I.; Muttukrishnan, R.; and Lloret, J. 2019. Human-Centric AI for Trustworthy IoT Systems With Explainable Multilayer Perceptrons. *IEEE Access* 7: 125562–125574. ISSN 2169-3536. doi:10.1109/ACCESS.2019.2937521. URL https://ieeexplore.ieee.org/document/8813027/.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 51(5): 93:1–93:42. ISSN 0360-0300. doi:10.1145/3236009. URL https://doi.org/10.1145/3236009.

Gunning, D.; and Aha, D. W. 2019. DARPA's Explainable Artificial Intelligence Program. *AI Magazine* 40(2): 44–58.

Hoffman, R.; Miller, T.; Mueller, S. T.; Klein, G.; and Clancey, W. J. 2018. Explaining Explanation, Part 4: A Deep Dive on Deep Nets. *IEEE Intelligent Systems* 33(3): 87–95. ISSN 1541-1672, 1941-1294. doi:10.1109/MIS.2018.033001421. URL https://ieeexplore.ieee.org/document/8423529/.

Hoffman, R. R.; and Klein, G. 2017. Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intelligent Systems* 32(3): 68–73. ISSN 1541-1672. doi:10.1109/MIS.2017.54. URL http://ieeexplore.ieee.org/document/7933919/.

Hoffman, R. R.; Mueller, S. T.; and Klein, G. 2017. Explaining Explanation, Part 2: Empirical Foundations. *IEEE Intelligent Systems* 32(4): 78–86. ISSN 1541-1672. doi:10.1109/MIS.2017.3121544. URL http://ieeexplore.ieee.org/document/8012316/.

Holland, J. H.; Holyoak, K. J.; Nisbett, R. E.; Thagard, P. R.; and Smoliar, S. W. 1987. Induction: Processes of Inference, Learning, and Discovery. *IEEE Expert* 2(3): 92–93. ISSN 0885-9000, 2374-9407. doi:10.1109/MEX.1987.4307100. URL https://ieeexplore.ieee.org/document/4307100/.

Horne, Z.; Muradoglu, M.; and Cimpian, A. 2019. Explanation as a cognitive process. *Trends in cognitive Sciences* 23(3): 187–199.

Klein, G. 2018. Explaining Explanation, Part 3: The Causal Landscape. *IEEE Intelligent Systems* 33(2): 83–88. ISSN 1541-1672. doi:10.1109/MIS.2018.022441353. URL https://ieeexplore.ieee.org/document/8378482/.

Lyons, J. B.; Clark, M. A.; Wagner, A. R.; and Schuelke, M. J. 2017. Certifiable Trust in Autonomous Systems: Making the Intractable Tangible. *AI Magazine* 38(3): 37–49. ISSN 2371-9621, 0738-4602. doi:10.1609/aimag.v38i3.2717. URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/2717.

Magazzeni, D.; Smith, D.; Langley, P.; and Biundo, S., eds. 2018. *Proceedings of the 28th International Conference on Automated Planning and Scheduling*. Delft, The Netherlands.

Mayes, G. R. 2020. Theories of Explanation. URL https://www.iep.utm.edu/explanat/.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38. ISSN 00043702. doi:10.1016/j.artint.2018.07.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0004370218305988.

Miller, T.; Howe, P.; and Sonenberg, L. 2017. Explainable AI: Beware of Inmates Running the Asylum. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Mueller, S. T.; Hoffman, R. R.; Clancey, W.; Emrey, A.; and Klein, G. 2019. Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. Darpa XAI Literature Report.

Mueller, S. T.; and Klein, G. 2011. Improving User's Mental Models of Intelligent Software Tools. *IEEE Intelligent Systems* 26(2): 77–83. doi:10.1109/MIS.2011.32.

Newell, A. 1982. The Knowledge Level. *Artificial Intelligence* (18): 87–127.

Potochnik, A. 2011. Explanation and understanding: An alternative to Strevens' Depth. *European Journal for Philosophy of Science* 1(1): 29–38. ISSN 1879-4912, 1879-4920. doi:10.1007/s13194-010-0002-6. URL http://link.springer.com/10.1007/s13194-010-0002-6.

Robert R. Hoffman; Klein, G.; and Mueller, S. T. 2018. Explaining Explanation For "Explainable Ai". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62(1): 197–201. ISSN 1541-9312. doi:10.1177/1541931218621047. URL http://journals.sagepub.com/doi/10.1177/1541931218621047.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5): 206–215. ISSN 2522-5839. doi:10.1038/s42256-019-0048-x. URL http://www.nature.com/articles/s42256-019-0048-x.

Shortlife, E. H.; Davis, R.; Axline, S. G.; Buchanan, B. G.; Green, C.; and Cohen, S. N. 1975. Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System. *Computers and Biomedical Research* 8: 303–320.

Swartout, W.; and Moore, J. D. 1985. Explainable (and Maintainable) Expert Systems. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, volume 1. Los Angeles.

Tsakiridis, N. L.; Diamantopoulos, T.; Symeonidis, A. L.; Theocharis, J. B.; Iossifides, A.; Chatzimisios, P.; Pratos, G.; and Kouvas, D. 2020. Versatile Internet of Things for Agriculture: An eXplainable AI Approach. In Maglogiannis, I.; Iliadis, L.; and Pimenidis, E., eds., *Artificial Intelligence Applications and Innovations*, volume 584, 180–191. Cham: Springer International Publishing. ISBN 978-3-030-49185-7 978-3-030-49186-4. doi:10.1007/978-3-030-49186-4_16. URL http://link.springer.com/10.1007/978-3-030-49186-4_16. Series Title: IFIP Advances in Information and Communication Technology.

Tulli, S.; Correia, F.; Mascarenhas, S.; Gomes, S.; Melo, F. S.; and Paiva, A. 2019. Effects of Agents' Transparency on Teamwork. In Calvaresi, D.; Najjar, A.; Schumacher, M.; and Främling, K., eds., *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, volume 11763, 22–37. Cham: Springer International Publishing. ISBN 978-3-030-30390-7 978-3-030-30391-4. doi:10.1007/978-3-030-30391-4_2. URL http://link.springer.com/10.1007/978-3-030-30391-4_2. Series Title: Lecture Notes in Computer Science.

van der Schaar, M. 2020. Machine learning: from black boxes to white boxes. URL https://www.youtube.com/watch?time_continue=3&v=EVl5iMpX1cg.

Wortham, R. H.; Theodorou, A.; and Bryson, J. 2016. What Does the Robot Think? Transparency as aWorht Fundamental Design Requirement for Intelligent Systems. In Kambhampati, S., ed., *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. New York, NY, USA: AAAI Press.

Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; and Sun, J. 2018. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. *arXiv:1711.08184 [cs]* URL http://arxiv.org/abs/1711.08184. ArXiv: 1711.08184.