Making Things Explainable vs Explaining: Requirements and Challenges Under the GDPR

Making Things Explainable vs Explaining: Requirements and Challenges under the GDPR

Francesco Sovrano[1 [0000−0002−6285−1041]], Fabio Vitali[1 [0000−0002−7562−5203]], and Monica Palmirani[2 [0000−0002−8557−8084]]

[1] DISI, University of Bologna
[2] CIRSFID-AI, University of Bologna

Abstract. The European Union (EU) through the High-Level Expert Group on Artificial Intelligence (AI-HLEG) and the General Data Protection Regulation (GDPR) has recently posed an interesting challenge to the eXplainable AI (XAI) community, by demanding a more user-centred approach to explain Automated Decision-Making systems (ADMs). Looking at the relevant literature, XAI is currently focused on producing explainable software and explanations that generally follow an approach we could term *One-Size-Fits-All*, that is unable to meet a requirement of centring on user needs. One of the causes of this limit is the belief that *making things explainable* alone is enough to have *pragmatic explanations*. Thus, insisting on a clear separation between *explainabilty* (something that can be explained) and *explanations*, we point to explanatorY AI (YAI) as an alternative and more powerful approach to win the AI-HLEG challenge. YAI builds over XAI with the goal to collect and organize explainable information, articulating it into something we called user-centred explanatory discourses. Through the use of explanatory discourses/narratives we represent the problem of generating explanations for Automated Decision-Making systems (ADMs) into the identification of an appropriate path over an explanatory space, allowing explainees to interactively explore it and produce the explanation best suited to their needs.

Keywords: Trustworthy AI · explanatorY AI (YAI) · XAI · HCI

1 Introduction

The academic interest in Artificial Intelligence (AI) [11] has grown together with the attention of Countries and people towards the possibly disruptive effects of ADM [38] in industry and the public administration (e.g., COMPAS [13], or in Italy the case-law "Buona Scuola"[3]), effects that may affect the lives of billions of persons [20]. Therefore governments are starting to act towards the establishment of ground rules of behaviour from complex systems, for instance through the enactment of the European GDPR[4], which identifies *fairness*, *lawfulness*, and in particular *transparency* as basic principles for every data processing tools handling personal data; even identifying a new *right to*

[3] Cons. stato, sez. VI, sent. 8 aprile 2019, n. 2270, Cons. Stato, sez. VI, sent. del 13 dicembre 2019, n. 8472, Cons. Stato, sez. VI, sent. del 4 febbraio 2020, n. 881.
[4] Regulation (EU) 2016/679.

explanation for individuals whose legal status is affected by a solely-automated decision. As a result, several expert groups, including those acting for the European Commission, have started asking the AI industry to adopt ethics code of conducts as quickly as possible [8, 14], drawing a set of expectations to meet in order to guarantee a *right to explanation*. These expectations define the goal of explanations under the GDPR and thus describe the requirements for explanatory content. Many interpretations have been given of what qualifies an explanation in this context, but among them we mention the one by the AI-HLEG, for its relevance and prominence. The AI-HLEG was established in 2018, by the European Commission, with the explicit purpose of applying the principles of the GDPR specifically to AI software, and produced a list of fundamental ethical principles for *Trustworthy AI* tools that include *fairness* and *explicability*. The *explicability* principle, in particular, means to provide alternative measures in case of "black box" algorithms like "traceability, auditability and transparent communication on system capabilities", in order to respect the fundamental rights. So it is important to provide information about *how* the ADM works, *what* is the final decision, *why* the ADM provides such conclusion, *which* data are used for training the AI and for the concrete real case processing. *Explicability* concerns the *ex-post* processing but also the *ex-ante* informative communication. Most importantly, according to the AI-HLEG, explanations should be "adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher)" and more over it "highly dependent on the context" [17], putting individual's needs at the centre, in a challenging way.

Notwithstanding these quite recent efforts, understanding what constitutes an explanation is a long-standing open problem. In literature there are various efforts in this direction and a long history of debates and philosophical traditions, often rooted in Aristotle's works and those of other philosophers. Among the many models proposed over the last few centuries some are now considered fallacious, albeit historically useful (e.g. Hempel et al.'s one [16]), in favour of more pragmatic (user-centred) ones (e.g. Achinstein's [2]). Despite this, Hempel et al.'s theory and Salmon's *Causal Realism* are probably the most (implicitly) mentioned and adopted models for explanations in AI, raising the question of whether technology is really aligned to the understandings of regulators and society or it is just acting conveniently. In fact, most of the literature on AI and explanations (e.g. eXplainable AI [3]) is currently focused on one-size-fits-all approaches usually able to produce only one type of explanations, defined through causal lens. Additional literature is focused on argumentation theory [9] or on sub-symbolic methodologies [7] for providing a deductive or inductive explanation.

It appears that this focus on pursuing one-size-fits-all explanations in XAI is justified by convenient definitions framing an explanation as the product of an act of making things explainable rather than a pragmatic (user-centred) act of explaining based on explainability. In other terms, there is no clear distinction between *making things explainable* and actually *explaining*. The exceptions to this pattern seem to be still too rare to be representative of disciplines like XAI. In this paper we take a strong stand against the idea that static, one-size-fits-all approaches to explanation have a chance of being pragmatic, thus meeting the AI-HLEG guidelines, and we propose to adopt a strong logical separation between *explainability* and *explaining*. In fact, we argue that explaining to humans is *computationally irreducible* and one-size-fits-all approaches (in

the most generic scenario) may suffer the curse of dimensionality as soon as the complexity of the explanandum surpasses a fairly trivial threshold. For example, a complex big-enough *explainable* software can be super hard to *explain*, even to an expert, and the optimal (or even sufficient) explanation might change from expert to expert. In this specific example, an explainable software is necessary but not sufficient for explaining. This is why we first draw a clear separation between XAI and explanatorY AI (YAI), which refers to systems that (given a "traditional" XAI system) are actually able to produce a satisfactory explanation ready to be delivered to a human user interested in examining the complex working and output of the system. Subsequently, we propose a model for YAI shaped on *discursive explanations*. Discursive explanations give a strong background of principles and means to create an interactive explanatory system that is able to produce user-centred explanations, by providing an explanatory space that is amenable to exploration by the users in order to create the explanation that best suits each one's background, needs and objectives.

This paper is structured as follows. In Section 2 we provide an introduction to the GDPR and the *Right to Explanation*, and we also provide a brief summary of the AI-HLEG Guidelines for Trustworthy AI. In Section 3, taking off from the GDPR and the AI-HLEG guidelines, we give a motivation of why user-centred explanatory tools are a key ingredient for Trustworthy AI. In this section we discuss the most prominent XAI issues to this end and the problem of *computational irreducibility* in explanations. In Section 4 we give an high-level overview of a possible model of User-Centred Explanatory Tool, defining YAI as a Explanatory Discursive Process responsible to collect and structure explainable information articulating it into user-centred explanations. Finally, in section 5 we conclude with a brief recap, pointing to a proof of concept.

2 Background: the Right to Explanation

The General Data Protection Regulation (GDPR) is an important 2016 EU regulation on personal data protection and the connected freedoms and rights. Since the GDPR is technology-neutral, it does not directly refer to AI, but several provisions are highly relevant to the use of AI for Automated Decision-Making system (ADM). For instance [19]:

- Principle 1. (a) requires personal data processing to be fair, lawful, transparent, necessary and proportional (Articles 5).
- Article 12 defines the obligations to fulfil a transparent information, communication and the modalities for the exercise of the data subject's rights.
- Articles 13-14-15 give individuals the right to be informed of the existence of solely automated decision-making, meaningful information about the logic involved, and the significance and envisaged consequences for the individual.
- Article 22 gives individuals the right not to be subject to a solely automated decision producing legal or similarly significant effects.
- Article 22(3) obliges organizations to adopt suitable measures to safeguard individuals when using solely automated decisions, including the right to obtain human intervention, to express his or her view, and to contest the decision.

Art. 22 defines the right to claim of a human intervention when a completely Automated Decision-Making systems (ADMs) may affect the legal status of a citizen. Art. 22 includes also several exceptions that derogate "to be subject to a decision based solely on automated processing" when the legal basis are supported by contract, consent or law. These conditions significantly limit the potential applicability of the right to explanation. For this reason in case of contract or consent the art. 22, paragraph 3 introduces the "right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision". Here explanations seem to be provided only after decisions have been made (*ex-post* explanations), and are not a required precondition to protest decisions. This is not completely true: in arts. 13-14-15 there is the obligation to inform about the "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved (Recital 63), as well as the significance and the envisaged consequences of such processing for the data subject." (*ex-ante* explanations). This combination of articles make the right of explanation very articulated and composed of different stages. Additionally, the recent White Paper on Artificial Intelligence [10] emitted by the European Commission stressed the need to monitor and audit not only the Automated Decision-Making system (ADM) algorithms but also the data records used for training, developing, running, the AI systems in order to fight the opacity and to improve transparency. From a technical point of view, there are technology-specific information to consider in order to fully meet the explanation requirements of the GDPR, for a more detailed overview refer to [35]. The qualities of explanations are listed in different works [25], but the EU Parliament [31] lists the following as a good summary of the current state of the art: intelligibility, understandability, fidelity, accuracy, precision, level of detail, completeness, consistency.

Article 22 is open to several interpretations [36, 28, 29] about whether providing individualised explanations is mandatory or just a good practice. To this end, Recital 71 provides interpretative guidance of Article 22. Two items are missing in Article 22 relative to Recital 71: the provision of "specific information" and the "right to obtain an explanation of the decision reached after such assessment". The second omission in particular raises the issue of whether controllers are really required by law to provide an individualised explanation. This issue is partially tackled by the AI-HLEG guidelines (endorsed by the EU Commission), giving further reason to believe that there is the intention to prefer user-centred explanations as soon as the technology is mature enough to guarantee them. At contrary Recital 63 requires *ex-ante* that the data subject should have the right to know and obtain communication in particular with regard to "the logic involved in any automatic personal data processing". The AI-HLEG tries to extend the GDPR expectations, targeting AI and giving further guidelines: accessibility and universal design should be a requirement for Trustworthy AI, with user-centrality at the core. This idea of a user-centred explanatory process find its roots in philosophy, for example in:

- Ordinary Language Philosophy [1, 22]: the act of explanation as the illocutionary attempt to produce understanding in another by answering questions in a pragmatic way.
- Cognitive Science [18, 22]: explaining as a process of belief revision, etc..

3 Problem Statement

Some of the limits in the current generation of XAI approaches have already been identified and spelt out by existing literature:

- "XAI has produced algorithms to generate explanations as short rules, attribution or influence scores, prototype examples, partial dependence plots, etc. However, little justification is provided for choosing different explanation types or representations" [37].
- "Research on explanation is typically focused on the person (or system) producing the explanation. [...] Does the explainee understand the system, concepts, or knowledge?" [25].
- "Much of XAI research tended to use the researchers' intuition of what constitutes a good explanation. There exist vast and valuable bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations, which argues that people employ certain cognitive biases and social expectations to the explanation process." [23]
- "XAI systems are built for developers, not users." [24, 25]
- etc..

To summarize, despite several efforts (e.g. [23, 12]) to tackle these issues, we can notice a majority of XAI tools lacking:

1. A broader vision: XAI should not involve only computer science, but also philosophy, psychology, cognitive science, etc..
2. Focus on user-centrality.
3. A consistent approach to evaluate the quality of explanations.

We claim that the cause of these limits are in the misunderstanding that explainability is enough for explaining. Indeed, by insisting on a clear logical separation between explainable systems and actual explanations, we argue that XAI is necessary but not sufficient for Trustworthy AI. In fact, XAI seems to be currently focused on producing explainable software and explanations that generally follow only a One-Size-Fits-All approach, failing to meet the user-centrality requirements. In the most generic scenario, explanations following a One-Size-Fits-All approach (*OSFA explanations*) should be considered not user-centred, by construction. For example, static representations where all aspects of a fairly long and complex computation are described and explained are one-size-fits-all explanations.

OSFA explanations have intuitively at least two problems:

1. if they are small enough to be simple, then in a complex enough domain they would not be able to generate an explanation containing enough information to satisfy the explanation appetite of every user, as the quantity of details required for satisfying every user would be necessarily larger than any small explanation in a few words.
2. if they contain all the necessary information, in a complex-enough domain they would contain an enormous amount of content and users interested in a specific aspect of the explanation would need to look for it within the whole explanation in hundreds or thousands of explanatory items mostly irrelevant to their purposes.

OSFA explanations could be useful for simple domains, but the complexity of a domain is exactly what motivates the need for explanations. In other terms, usefulness of explanations is obviously greater in complex domains.

An interesting parallel, to show the second problem, is that of surveillance cameras in front of a bank door. Surveillance cameras continuously record and make available to the investigators hundreds and hundreds of hours of excellent quality videos that allow the precise identification of thousands of people passing under the cameras. But our investigator is not interested in hundreds of hours of video, but only in those three seconds in which a suspect person in need to be identified was under the cameras. The relevance of these few seconds (out of hundreds of hours) is entirely based on the specific investigative task, which depends on the function that the investigator gives to the identification of the person, and this function depends on the purpose of identification (i.e. Is he the robber? A possible accomplice? A witness?). The purpose of the investigation is known to the investigator but not to the surveillance system, and in many cases it cannot be decided in advance but it becomes clear only during the evolution of the investigation. Similarly, the interest of a user in the output of an explanation system often may lie on a few short statements out of the hundreds of thousands that the explanation system may be able to generate, and these few ones depend on the function that the user gives to the explanation. This is why we must assume that in general the purpose of the explanation is known to the user but not to the explanation system, and it cannot be decided in advance but it becomes clear only during the evolution of the task in which the explanation is required. This phenomenon is known also as *computational irreducibility* [39] and it is typical of emerging phenomena, such as physical, biological and social ones [5].

A user-centred explanatory tool requires to provide goal-oriented explanations. Goal-oriented explanations implies explaining facts that are relevant to the user, according to her/his background knowledge, interests and other peculiarities that make her/him a unique entity with unique needs that may change over time. The computational irreducibility issue raises the following questions:

1. How to model and create a *user-centred* explanatory process, without rewriting the tool for every different user?
2. How to evaluate the quality of an explanatory process?

4 Proposed Solution

In order to answer the first question we propose to:

- Disentangle *explainability* from *explaining*: that is separate the presentation logic (*explaining*) from the application logic (*explainability*). In fact, only *explaining* has to be user-centred.
- Design a presentation logic that would allow personalised explanations given the same explainable information.

In figure 1 we show a simple model of an Explanatory Tool for Trustworthy AI, obtained by our own need to clearly separate between explainabilty and explanations.

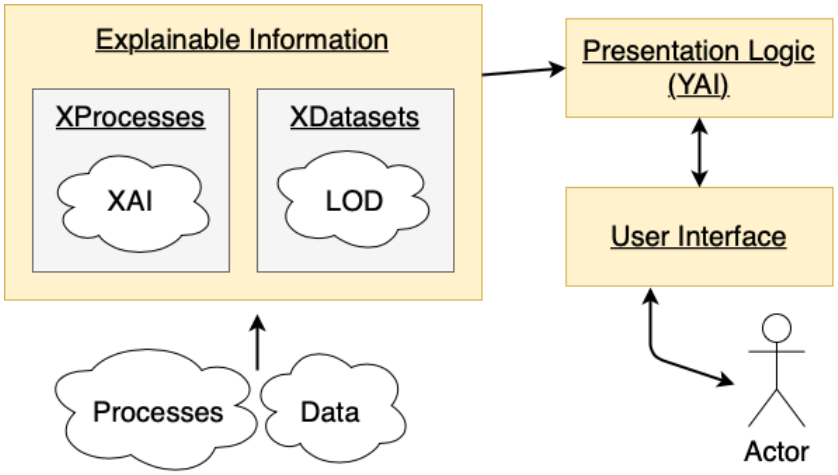

Fig. 1. XAI vs YAI: an abstract model of Explanatory Tool for Trustworthy AI. This model shows how to decompose the flow of explanatory information that moves from raw representations of processes/data to the explainee (or actor). Raw data are refined into explainable datasets - e.g. Linked Open Data (LOD), etc.. Raw processes are refined into explainable processes. Explainable information can be used by YAI to generate pragmatic explanations.

More in detail, to increase the overall cohesion of the system, in this model we require an explicit logical separation between the functionalities related to *producing explainable information*, and those related to *producing pragmatic explanations*. In addition, we envision another logical separation in the production of actual explanations between *building explanations* (i.e. the presentation logic) and *interfacing with users*. Independently, producing explainable information should be separated in *generating explainable processes* and *producing explainable data-sets*. Thus, the main modules involved in the model are:

- The Explainable Information (EI) module, made of the eXplainable Processes (XP) and the eXplainable Datasets (XD) sub-modules.
- The YAI or Presentation Logic module.
- The User Interface (UI) module.

In other terms, we propose to distinguish between eXplainable AI (XAI) and explanatorY AI (YAI), considering them as different components of Trustworthy AI. We like to say that Trustworthy AI needs both the Xs and the Ys of AI[5].

The YAI module is the module responsible to collect and structure explainable information articulating it into user-centred explanations. In other terms, defining the YAI module is the same of defining a *user-centred explanatory process*. We are interested in defining a user-centred explanatory process aligned to the GDPR and the AI-HLEG guidelines. Speaking of user-centrality, we may assume that different types/groups of

[5] XX and XY are human chromosomes responsible for gender.

users exist: lay person, expert, legal operators, etc.. each one with its own background knowledge and unique characteristics. If the explanations have to be tailored, does this imply that we should have a different explanatory tool for every possible different user? Probably not. We believe that an explanatory tool is an instrument for articulating explainable information into an *explanatory discourse*. This definition of explanatory tool is drawn from the essential best-practices of scientific inquiry, involving [6]:

- Sense-making of phenomena: classical question answering to collect enough information for understanding, thus building an explainable explanandum (perhaps through XAI).
- Articulating understandings into discourses: re-ordering and aggregation of explainable information to form an explanatory narrative or more generally a discourse to answer research questions.
- Evaluating: pose and answer questions about the quality of the presented information; e.g., argument them in a public debate.

Therefore we define a user-centred explanatory discourse as: "A sequence of information (explanans) to increase understanding over explainable data and processes (explanandum), for the satisfaction of a specified explainee that efficiently and effectively interacts with the explanandum (interaction) having specific goals in a specified context of use". Our definition takes inspiration from [32, 26, 21], integrating concepts of usability defined in ISO 9241 (Ergonomics of Human System Interaction [15]), such as the insistence on the term "specific", the triad "explainee", "goal" and "context of use", as much as the identification of specific quality metrics, which in our case are *effectiveness*, *efficiency* and *satisfaction*.

Similarly to how *satisfaction* has increased in importance in user experience studies in recent years, we believe that satisfaction should be considered one of the most important metrics for the assessment of the quality of explanations, too. The qualities of the explanation that provides the explainee with the necessary *satisfaction*, using the categories provided by [26], can be summarized in a good choice of narrative appetite, structure and purpose. To understand "narrative appetite" we have to consider that "in order for a narrative discourse to flourish, both parties (the narrator and the reader) have to find engagement in this social transaction interesting enough to prevail over competing activities. Thus, stories must not only be accounts of events, but accounts of events that someone cares to know more about; we must want to know what happened if we are to continue reading or listening." This appetite can be quenched by the proper structuring of the narration: "Narrative, we have shown, is a narrator's recounting of *events structured in time*. The elements of both time and structure are associated in many descriptions of narrative". In addition, "The element of *connectability* [...] structures different texts. Connectability [...] must be strictly observed in expository texts where an argument is to be developed or information is to be conveyed. In such texts, the writer aims for a precise interpretation where a multiplicity of possible meanings must be constantly narrowed down". Finally, the identification of purpose in narratives is central: "stories are constructed to help us understand the world we live in: to help comprehend the life that is in me and around me. [...] it is through narrative that we are able to accommodate the new within that which is familiar to us. In these descriptions

of purpose, narrative can be interpreted as helping us better understand the natural as well as the human world".

The problems of a user-centred approach to explanations is that fully-automated explanatory processes are unlikely to target quality parameters that guarantee the satisfaction of all specified explainees, as described above, due to the *computational irreducibility* of the process of explaining. Even if an AI could be used to generate such user-centred explanations, in the context of explanations under the GDPR this would only shift the problem of explaining from the original ADM to another ADM (the explanatory AI that explains the original ADM). As such we believe that (at least for the explanations under the GDPR), the most straightforward solution is to encourage readers (explainees) and narrators (explainers) to become one, *users* generating the narration for themselves by selecting and organizing narratives of individual event-tokens according to the structure that best caters their appetite and purpose. In this sense, a tool for creating explanatory discourses would allow users to build intelligible sequences of information, containing arguments that support or attack the claims underlying the goal of an *explanatory narrative process*. This idea of data controllers and data subjects "becoming one" can be understood in a twofold way. First, at its best possible light, such tool should convince and dissuade data subjects to ask for human intervention, e.g. Art 22(3) of the GDPR. Second, the tool should help data controllers to abide by the law, by illustrating the decision that can be contested by data subjects.

An *explanatory narrative* is always only one of the many possible narratives that can be built to shed light on an explanandum. All the possible narratives for an explanandum form a complex network of information that we call *Explanatory Space*. In this sense, an explanatory discourse is a path within an Explanatory Space. As analogy, we might see the Explanatory Space as a sort of manifold space where every point within it is interconnected information about one or more aspects of the explanandum. So that every point of the Explanatory Space is not user-centred locally, but globally as an element of a sequence of information that can be chosen by a user according to its interest drift while exploring the space.

As mentioned in Section 3, the amount of information forming such Explanatory Spaces can be overwhelming, given any complex-enough explanandum. Thus, in order to answer our research question, what we need is to design a process to effectively allow users to extract explanatory narratives from an Explanatory Space. In [35] we present our model of Explanatory Narrative Process making specific references to the GDPR and the AI-HLEG guidelines, modelling a generic explanatory process, giving a formal definition of explanandum, explanans and Explanatory Space. Hereafter we show a plausible example of YAI in action.

4.1 Example

Let's consider the following example where a user-centred explanatory tool is used to explain the decision taken by an ADM on a case concerning the GDPR, art. 8. The aforementioned case is about the conditions applicable to child's consent in relation to information society services. The art. 8 of GDPR fixes at 16 years old the maximum age for giving the consent without the parent-holder authorization. This limit could be derogated by the domestic law. In Italy the legislative decree 101/2018 defines this limit

at 14 years. In this situation we could model legal rules in LegalRuleML [4, 30] using defeasible logic, in order to be able to represent that the GDPR art. 8 rule (16 yearsOld) is overridden by the Italian's (14 yearsOld). The SPINDle legal reasoner processes the correct rule according to the jurisdiction (e.g., Italy) and the age. Suppose that Marco (a 14 years old Italian teenager living in Italy) uses Whatsapp, and his father, Giulio, wants to remove Marco's subscription to Whatsapp because he is worried about the privacy of Marco when online. In this simple scenario, the Automated Decision-Making system (ADM) system would reject Giulio's request to remove Marco's profile, because of the Italian legislative decree 101/2018. What if Giulio wants to know the reasons why his request was rejected? Figure 2 shows a possible view of a user-centred explanatory tool based on our model. Thanks to the user-centred explanatory tool Giulio can actually choose what information to expand and consider, building its own personalised explanatory discourse out of a predefined Explanatory Space.

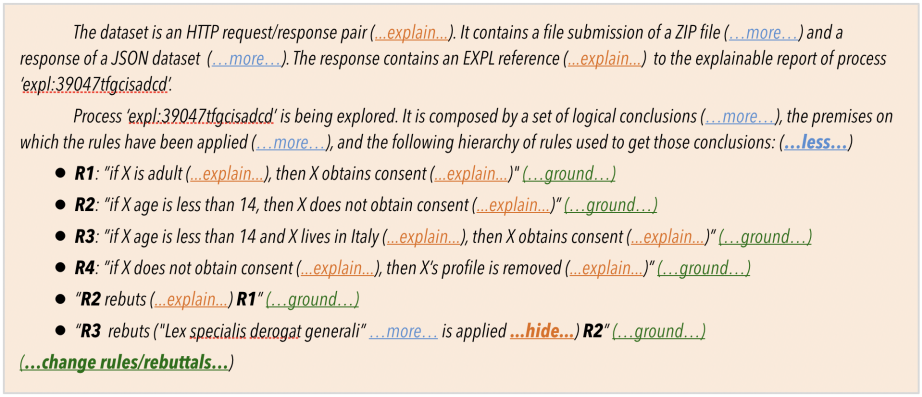

Fig. 2. Example of explainer: underlined coloured words represent different possible actions a user can operate to explore the Explanatory Space, extracting its own narrative. For example, clicking on a "...more..." button the user can expand the explanans.

5 Conclusions

In this paper we analysed some of the limits in the current generation of XAI approaches, with respect to the goals of Trustworthy AI set by the GDPR and the AI-HLEG guidelines, identifying the cause of these limits in the misunderstanding that *making things explainable* is enough for *pragmatically explaining*. Indeed, by insisting on a clear logical separation between explainable systems and actual explanations, we argued that XAI is necessary but not sufficient for Trustworthy AI, therefore presenting an abstract model of explanatorY AI (YAI). In our model, YAI builds over XAI and it is intended to be a set of tools for organising the presentation logic of a user-centred explanatory software in a way that would allow personalised explanations about complex-enough explananda by generating *discursive explanations* out of an Explanatory Space.

In this paper we take a strong stand against the idea that static, one-size-fits-all approaches to explanation have a chance of being pragmatic, thus meeting the AI-HLEG guidelines. For a concrete proof of concept of YAI (including software and experiment analysis) we point the reader to our most recent works, e.g. [34].

Finally, it is clear that the solution we proposed avoids the problem which relates to balancing between what is possible in terms of formal explainability and what is required as to the level of detail of information regarding the "logic of processing". In other words, we assumed that systems in question can be both formally explainable and pragmatically able to be explained. So, we leave as future work an analysis of what are the minimum requirements for information to be considered explainable enough for pragmatic explanations with a proper degree of exactness, detail and fruitfulness[6]. This might help also to perform a reasonable impact assessment of the ADM, as defined by art. 35 of the GDPR.

Acknowledgements

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the MSCA grant agreement No 690974 "MIREL: MIning and REasoning with Legal texts". Last but not least, a big thank you to all the reviewers for their brilliant comments and different insights.

[6] Perhaps drawing from Carnap's theory [27].

Bibliography

[1] Peter Achinstein. *The nature of explanation*. Oxford University Press on Demand, 1983.

[2] Peter Achinstein. *Evidence, explanation, and realism: Essays in philosophy of science*. Oxford University Press, 2010.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[4] Tara Athan, Harold Boley, Guido Governatori, Monica Palmirani, Adrian Paschke, and Adam Z Wyner. Oasis legalruleml. In *ICAIL*, volume 13, pages 3–12, 2013.

[5] Brian Beckage, Stuart Kauffman, Louis J Gross, Asim Zia, and Christopher Koliba. More complex complexity: Exploring the nature of computational irreducibility across physical, biological, and human social systems. In *Irreducibility and computational equivalence*, pages 79–88. Springer, 2013.

[6] Leema Kuhn Berland and Brian J Reiser. Making sense of argumentation and explanation. *Science Education*, 93(1):26–55, 2009.

[7] Omicini A. Calegari R., Ciatto G. On the integration of symbolic and sub-symbolic techniques for xai: A survey. *Intelligenza Artificiale*, 2020.

[8] Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. Artificial intelligence and the 'good society': the us, eu, and uk approach. *Science and engineering ethics*, 24(2):505–528, 2018.

[9] Toni F. Cocarascu O., Rago A. Explanation via machine arguing. In *Reasoning Web. Declarative Artificial Intelligence. Reasoning Web 2020*, pages 53–84. Springer, 2020.

[10] EU Commission. White paper - on artificial intelligence - a european approach to excellence and trust, com(2020) 65 final, 2020.

[11] European Commission. *COM(2018) 237 final Brussels, Artificial Intelligence for Europe*. European Commission, 2018.

[12] DARPA. Broad agency announcement explainable artificial intelligence (xai). *DARPA-BAA-16-53*, pages 7–8, 2016.

[13] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[14] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28 (4):689–707, 2018.

[15] International Organization for Standardization. *Ergonomics of human-system interaction: Part 210: Human-centred design for interactive systems*. ISO, 2010.

[16] Carl G Hempel et al. Aspects of scientific explanation. 1965.
[17] AI HLEG. Ethics guidelines for trustworthy ai, 2019.
[18] John H Holland, Keith J Holyoak, Richard E Nisbett, and Paul R Thagard. *Induction: Processes of inference, learning, and discovery*. MIT press, 1989.
[19] ICO. Project explain interim report. https://ico.org.uk/about-the-ico/research-and-reports/project-explain-interim-report/, 2019. Online; accessed 05-Jan-2020.
[20] Richard Kuhn and Raghu Kacker. An application of combinatorial methods for explainability in artificial intelligence and machine learning (draft). Technical report, National Institute of Standards and Technology, 2019.
[21] Peter Lipton. What good is an explanation? In *Explanation*, pages 43–59. Springer, 2001.
[22] GR Mayes. Theories of explanation. the internet encyclopedia of philosophy, 2005.
[23] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.
[24] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.
[25] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*, 2019.
[26] Stephen P Norris, Sandra M Guilbert, Martha L Smith, Shahram Hakimelahi, and Linda M Phillips. A theoretical framework for narrative explanation in science. *Science Education*, 89(4):535–563, 2005.
[27] Catarina Dutilh Novaes and Erich Reck. Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. *Synthese*, 194(1):195–215, 2017.
[28] Ugo Pagallo. Algoritmi e conoscibilità. *Rivista di filosofia del diritto*, 2020.
[29] Monica Palmirani. Big data e conoscenza. *Rivista di filosofia del diritto*, 2020.
[30] Monica Palmirani and Guido Governatori. Modelling legal knowledge for gdpr compliance checking. In *JURIX*, pages 101–110, 2018.
[31] EU Parliament. Understanding algorithmic decision-making: Opportunities and challenges, 2019.
[32] John Passmore. Explanation in everyday life, in science, and in history. *History and Theory*, 2(2):105–123, 1962.
[33] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.
[34] Francesco Sovrano and Fabio Vitali. From philosophy to interfaces: an explanatory method and a tool based on achinstein's theory of explanation. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 2021.
[35] Francesco Sovrano, Fabio Vitali, and Monica Palmirani. Modelling gdpr-compliant explanations for trustworthy ai. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 219–233. Springer, 2020.

[36] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 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, 2017.

[37] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 601. ACM, 2019.

[38] WP29. *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01)*. European Commission, 2016.

[39] Hervé Zwirn and Jean-Paul Delahaye. Unpredictability and computational irreducibility. In *Irreducibility and Computational Equivalence*, pages 273–295. Springer, 2013.