Modelling GDPR-Compliant Explanations for Trustworthy AI

(Article begins on next page)

20 April 2024

Modelling GDPR-Compliant Explanations for Trustworthy AI

Francesco Sovrano[1], Fabio Vitali[1], and Monica Palmirani[2]

[1] DISI, University of Bologna
[2] CIRSFID, University of Bologna
{francesco.sovrano2, fabio.vitali, monica.palmirani}@unibo.it

Abstract. Through the General Data Protection Regulation (GDPR), the European Union has set out its vision for Automated Decision-Making (ADM) and AI, which must be reliable and human-centred. In particular we are interested on the Right to Explanation, that requires industry to produce explanations of ADM. The High-Level Expert Group on Artificial Intelligence (AI-HLEG), set up to support the implementation of this vision, has produced guidelines discussing the types of explanations that are appropriate for user-centred (interactive) Explanatory Tools. In this paper we propose our version of Explanatory Narratives (EN), based on user-centred concepts drawn from ISO 9241, as a model for user-centred explanations aligned with the GDPR and the AI-HLEG guidelines. Through the use of ENs we convert the problem of generating explanations for ADM 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. To this end we list suitable exploration heuristics, we study the properties and structure of explanations, and discuss the proposed model identifying its weaknesses and strengths.

Keywords: Interactive Explanatory Tool · General Data Protection Regulation · Trustworthy Artificial Intelligence.

1 Introduction

The academic interest in Artificial Intelligence (AI) has grown together with the attention of countries and people toward the actual disruptive effects of Automated Decision Making (ADM [27]) in industry and in the public administration, effects that may affect the lives of billions of people. Thus, GDPR (General Data Protection Regulation, UE 2016/679) stresses the importance of the Right to Explanation, with 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 [5,9]. The GDPR draws a set of expectations to meet in order to guarantee the Right to Explanation (for more details see section 2). These expectations define the goal of explanations under the GDPR and thus describe requirements for explanatory content that should

be "adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher)" [11]. Analysing these requirements we found a minimal set of explanation types that are necessary to meet these expectations: causal, justificatory and descriptive.

Most of the literature on AI and explanations (e.g. eXplainable AI) is currently focused on one-size-fits-all approaches usually able to produce only one of the required explanation types: causal explanations. In this paper we take a strong stand against the idea that static, one-size-fits-all approaches to explanation (explainability) have a chance of satisfying GDPR expectations and requirements. In fact, we argue that one-size-fits-all approaches (in the most generic scenario) may suffer the curse of dimensionality. 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. This is why we argue that an explanatory tool for complex data and processes has to be user-centred and thus interactive. The fact that every different user might require a different explanation does not imply that there might be no unique and sound process for constructing user-centred explanations. In fact we argue that every interactive explanatory tool defines and eventually generates an Explanatory Space (ES) and that an explanation is always a path in the ES. The explainee explores the ES, tracing a path in it, thus producing its own (user-centred) explanation.

This is why we assert that a more nuanced approach must be considered, where explanations are user-centred narratives that allow explainees to increase understanding through sense-making and articulation, in a manner that is fit to specified explainees, their goals, and their context of use. Upon these considerations, and following the High-Level Expert Group on Artificial Intelligence (AI-HLEG) Ethics Guidelines for Trustworthy AI, we propose here a new model of a User-Centred Explanatory Tool for Trustworthy AI, compliant with GDPR. To this extent we propose a definition of user-centred explanations as Explanatory Narratives, based on concepts drawn from ISO 9241. We present a formal model of an Interactive Explanatory Process consequently identifying 4 fundamental properties (the SAGE properties) of a good explanation and 3 heuristics for the exploration of the ES. Finally, we define the structure of the ES, combining the SAGE properties and the identified exploration heuristics, thus showing an application of the model to a real-case scenario.

This paper is structured as follows: in Section 2 we provide some background information. In Section 3 we discuss over the GDPR, introducing our definition of Explanatory Narratives. In Section 4 we analyse existing work. While in Section 5 we propose a simple model of a User-Centred Narrative Explanatory Process. Finally in Section 6 we discuss the strengths and the weaknesses of our model, and in Section 7 we conclude pointing to future work.

2   Background

2.1   Explanations in Literature

In literature, many types of possible explanations have been thoroughly discussed, and it is not clear whether a complete and detailed taxonomy may exist. In the field of Explainable Artificial Intelligence (XAI), the most discussed type of explanations is probably the causal one. We can say that explanations can be causal or non-causal. Causal explanations may have many different shapes and flavours [16], including explanations based on *causal attributions* (or chains), on *causal reasoning*, etc.. Similarly, non-causal explanations can be of several different types, including (but not limited to):

- Descriptive: explanations related to conceptual properties and characteristics: hypernyms, hyponyms, holonyms, meronyms, etc.
- Justificatory: explanations of why a decision is good.
- Deontic: justifications of the decision based on permissions, obligations and prohibitions [15]. In this sense, deontic explanations are a subset of justificatory explanations.
- Contrastive: counterfactual explanations on events instead of the causes of events.

2.2   The GDPR and the Right to Explanation

The GDPR is technology-neutral, so it does not directly reference AI, but several provisions are highly relevant to the use of AI for decision-making [13]. The GDPR defines the "Right to Explanation" as a right that individuals might exercise when their legal status is affected by a solely automated decision. In order to put the user in the conditions to be able to contest an automated decision and thus to exercise the right to explanation, the insights of the decisions have to be properly explained.

The GDPR defines (indirectly) two modalities of explanation: explanations can be offered before (*ex-ante*; artt. 13-14-15) or after decisions have been made (*ex-post*; art. 22, paragraph 3). For each modality, the GDPR defines goals and purposes of explanations, thus providing a set of explanatory contents. From a technical point of view, there are technology-specific information to consider in order to fully meet the GDPR explanation requirements. Fundamentally, ex-ante we should provide information that guarantees the transparency principle, such as describing:

- The algorithms and models pipeline composing the ADM.
- The data used for training (if any), developing and testing the ADM.
- The background information (e.g. the jurisdiction of the ADM).
- The possible consequences of the ADM on the specific data subject.

Ex-post the data subject should be able to fruitfully contest a decision, so he/she should be given access to:

- The justification about the final decision.
- The run-time logic flow (causal chain) of the process determining the decision.
- The data used for inferring.
- Information (metadata) about the physical and virtual context in which the automated process happened.

In this scenario, law and ethics scholars have been more concerned with understanding the internal logic of decisions as a means to assess their lawfulness (e.g. prevent discriminatory outcomes), contest them, increase accountability generally, and clarify liability. For example, [26] propose counterfactuals as a reasonable way to lawfully provide *causal* explanations under the GDPR's right to explanation. [26] takes strength from the Causal Inference theory, in which counterfactuals are hypothesised to be one of the main tools for Causal Reasoning [20].

2.3   Transparency and the AI-HLEG Guidelines for Trustworthy AI

The AI-HLEG has been charged by the European Union to identify a set of Ethics Guidelines for Trustworthy AI, published in April 2019 [11]. The AI-HLEG vision for a user-centred AI appears to incorporate the GDPR principles, trying to expand them into a broader framework based on 4 consolidated ethical principles, including: respect for human autonomy, fairness, explicability. From the aforementioned principles they derive seven key requirements for Trustworthy AI, including: transparency, diversity, non-discrimination and fairness, accountability.

The ethical principle of Explicability [9] is associated to the requirements of Transparency and Accountability. The Transparency requirement, in turn, is clearly inspired by articles 13-14-15-22 of the GDPR. In a way, the AI-HLEG applies the technologically neutral GDPR by defining relevant guidelines on how Transparency can be achieved in Trustworthy AI systems, also through accessibility and universal design. The "Accessibility and Universal Design" requirement puts user-centrality at the core of Trustworthy AI systems. While the Transparency requirement encompasses transparency of elements relevant to an AI system (the data, the system, the business models), including: Traceability, Explainability, and Communication

3   GDPR-compliant interactive explanatory tools for Trustworthy AI

In this section we will discuss over the GDPR and the AI-HLEG guidelines, identifying a minimal set of required explanation types, thus proposing a new User-Centred Explanatory Tool based upon a definition of Explanatory Narratives aligned to ISO 9241.

3.1 Explanations under GDPR

The GDPR clearly draws a set of expectations to meet, in order to guarantee the Right to Explanation. These expectations are meant to define the goal of explanations and thus an explanatory content that may evolve together with technology. This explanatory content identifies at least 3 different types of explanations: causal, descriptive, justificatory. We will refer to them as the minimal set of explanations required, for explaining ADM under the GDPR.In fact, in the case of GDPR, we see that:

- Descriptive explanations are mostly required in the *ex-ante* phase, to explain business-models, the possible effects of ADM on user, and characteristics and limitations of the algorithms.
- Causal explanations are mostly required in the *ex-post* phase, to explain the causes of a solely automated decision.
- Justificatory explanations are required in both the *ex-ante* and *ex-post* phase, to justify decisions through permissions, obligations and so on.

The aforementioned explanations can be provided to the user through one or more explanatory tools as part of the whole AI system. This is why the AI-HLEG has defined some characteristics that these AI systems (and consequently their explanatory tools) should possess for trustworthiness. These characteristics include (among other things) transparency and user-centrality.

3.2 User-Centred Explanatory Tools

According to the AI-HLEG and the ICO [13], user-centrality implies that (in the most generic scenario) explanations following a One-Size-Fits-All approach (OSFA explanations) are not user-centred by design. For example, static symbolic 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. The first problem is that if they are small-enough to be simple, it is impossible that in a complex-enough domain they would contain enough information to satisfy the explanation appetite of every user. The second problem is that if they contain all the necessary information, in a complex-enough domain they would contain an enormous amount of information and every user interested in a specific fragment of the explanation might look for it within hundreds of pages of explanations mostly irrelevant to her/his purposes.

At this point one might observe that OSFA explanations could be useful for simple domains, but according to [22] the complexity of a domain is exactly what motivates the need for explanations. In other terms, explanations are more useful to be given in complex domains.

What are OSFA explanations? Static explanations are OSFA explanations by design, but sometimes OSFA explanations can also be interactive. In fact, intuitively, simply adding naive ways of interaction to a static explanatory tool

does not imply that the new interactive tool is no more following a one-size-fits-all approach. This is why we argue that interactivity is not a sufficient property for user-centred explanatory tools.

What are the sufficient properties for user-centred explanatory tools? A user-centric explanatory tool requires to provide goal-oriented explanations. Goal-oriented explanations implies explaining facts that relevant to the user, according to its background knowledge, interests and other peculiarities that make her/him a unique entity with unique needs that may change over time. If the explanations have to be adapted to users, does this imply that we should have a different explanatory tool for every possible user?

3.3   Explanatory Narratives

The fact that every different user might require a different explanation does not imply that there might be no unique and sound process for constructing user-centred explanations. In fact we argue that every interactive explanatory tool defines and eventually generates an Explanatory Space (ES) and that an explanation is always a path in the ES. The explainee explores the ES, tracing a path in it, thus producing its own (user-centred) explanation. We are going to give a more formal definition of the ES and the other components of an explanatory process, later.

Actually, being able to construct useful explanations is one of the main challenges of making science. This is why a lot of literature exist on how to construct scientific explanations. Constructing scientific explanations and participating in argumentative discourse are seen as essential practices of scientific inquiry (e.g., [8]), that according to [4] involves 3 different practices: sense-making, articulating, evaluating. In fact a scientist should use evidence to make sense of phenomenon, articulating understandings into explanatory narratives. These explanatory narratives should be validated, e.g., defending them in a public debate against the attacks of scientific community. We believe that similarly to scientific explanations, constructing lawful explanations involves the same practices. For example, legal evidential reasoning can be seen as reasoning on evidences (sense-making) in order to justify/prove an hypothesis (articulating), in a way that the resulting arguments can be defended from opponents and accepted by judges during a debate (evaluating) [21]. This is why we argue that a user-centred explanatory tool should be an instrument for sense-making, articulating and evaluating information into an explanatory narrative. If we focus on user-centred explanatory tools, then we are focusing on tools for sense-making through creating an explanatory narrative. What is an explanatory narrative? We consider an explanatory narrative as a sequence of information (explanans) to increase understanding over explainable data and processes (explanandum), for the satisfaction of a specified explainee that interacts with the explanandum having specified goals in a specified context of use. Our definition takes inspiration from [19, 17, 14], integrating concepts of usability defined in ISO 9241, such as the insistence on the term "specific", the triad "explainee", "goal" and "context of use", as much as the identification of a specific quality metric, which in our case

are effectiveness and satisfaction. The qualities of the explanation that provide the explainee with the necessary satisfaction, following the categories provided by [17], can be summarized in a good choice of narrative appetite, structure and purpose.

The problems of a user-centred approach is that fully-automated explainers are unlikely to target quality parameters that guarantee the satisfaction of each specified explainee. Even if an AI could be used to generate such user-centred explanations, this would only shift the problem of explaining from the original ADM to another ADM – the explanatory AI used to explain the original ADM. As such we believe that, in the case of user-centred explanations, the simplest solution is to require that reader (explainee) and narrator (explainer) are the same, 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 narratives would allow users to build intelligible sequences of information, containing arguments that support or attack the claims underlying the goal of a narrative explanatory process defined by explainee/explainer and explanandum.

4   Related work

Apparently most of the tools for explaining ADM are static (e.g. AIX360 [1]). Interactive tools also exist [10, 6, 7, 28], but: 1) they do not consider to offer descriptive explanations, but other types of explanations on pre-defined aspects of the ADM; 2) or they generate explanations automatically (e.g from static argumentation frameworks), using templates. Completely automated sense-making or understanding articulation is possible only with very specific ADMs, or by pre-defining narrative scenarios that can be as powerful as dangerous [3, 25]. Furthermore, a solely automated explainer is an ADM process itself, that might require to be explained as well.

Actually, as far as we know, there is no tool for explanatory narratives of ADMs, but more generic tools for teaching exist such as [24, 23]. There are some interesting similarities between our work and the two aforementioned works, including the assumption that highlighting the structural elements of the explanandum is necessary in the articulation of an understanding.

5   A User-Centred Narrative Explanatory Process

In this section we present our model of User-Centred Narrative Explanatory Process under the GDPR. We start modelling a generic explanatory process, giving a formal definition of explanandum, explanans and Explanatory Space. Concurrently to modelling we show, step-by-step, an application of the model in a real-case scenario.

5.1   Real-Case Scenario

We present here a real-case scenario we will continuously refer to when defining our user-centred narrative explanatory process. This real-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 [2, 18] using defeasible logic, in order to be able to represent the fact that the GDPR art. 8 rule is overridden with the Italian's one. 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 ADM system would reject Giulio's request to remove Marco's profile.What if Giulio wants to get an explanation of the automated decision? To answer this question we have to pick an explanatory process.

5.2   The Interactive Explanatory Process
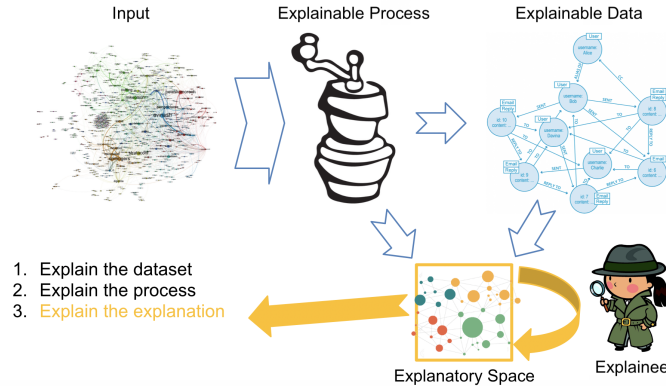
Fig. 1. Stylized interactive explanatory process.

According to the definition given in Section 3.3, a user-centred (interactive) narrative explanatory process is a process explaining an explanandum to an explainee (reader and narrator), thus producing as output an explanans (explanatory narrative) that is meaningful for the specific explainee. As shown in figure 1, an explanatory process is a function p for which $p(D, E_t, i_t) = E_{t+1}$, where:

- D is the explanandum.
- E_t is the input explanans, at step $t \geq 0$.

- i_t is the user interaction at step t.
- E_{t+1} is the output explanans, thus at step $t + 1$.

We can iteratively apply p, starting from an initial explanans E_0, until satisfaction. The user interaction i is a tuple made of an action a taken from the set A of possible actions, and a set of auxiliary inputs required by the action a.

5.3   The Explanandum

As defined by AI-HLEG guidelines, in this setting the explanandum is a collection of context-dependent information, made of one or more: datasets, processes, business models (that are higher-order processes). Naturally, the explanandum has to be *explainable* in order to be explained. In this scenario, datasets and processes are said to be explainable when they have a clear and not ambiguous symbolic representation, i.e., when their data items and rules are aligned with meaningful ontologies[3] for the end-users. The context-dependency implies that, in the most generic scenario, the information contained in datasets and processes is not sufficient to understand their nature without external knowledge. This external knowledge is commonly assumed to be part of the explainee knowledge (e.g. the knowledge about how to interpret a natural language). We will refer to this external common knowledge as the *explanandum context*, considering it as a dataset implicitly part of the explanandum.

In our real-case the explanandum is made of: a rule-base (the LegalRuleML one), a dataset of premises (the information about the involved entities), a dataset of conclusions obtained by applying the premises to the rule-base, a causal chain (the ordered chain of rules involved in the production of the dataset of conclusions).

What are datasets and processes?
A dataset is a tuple $\langle X, N, S \rangle$ where:

- X is a set of data-items.
- N is the set of (possibly informal) ontologies describing X.
- S are the (possibly unstructured) sources used to derive N and X.

A process is a tuple $\langle D_1, D_2, D_3 \rangle$ where:

- D_1 is the dataset of the process *inputs*, i.e. the domain.
- D_2 is the dataset of the process *function*.
- D_3 is the dataset of the process *outputs*, i.e. the codomain.

When D_1 is a collection of processes, the process is said to be higher-order.

In our real-case the sources of the function/rule-base are the GDPR (art. 8) and the Italian legislative decree 101/2018, while the source of the inputs/premises and the outputs/conclusions is the textual description of the case. The main ontology of the rule-base is a representation of the knowledge behind the SPINDle legal reasoner, while the ontologies for the premises and the conclusion might include the Dublin Core Schema, etc..

---

[3] These ontologies do not have necessarily to be explicit, formal or complete.

5.4   The Explanans and The Explanatory Space

The explanans is a particular type of dataset, in fact it is an ordered sequence of arguments (contained in the explanandum) useful to the explainee to reach its goals. The set of all the possible explanans reachable by an explainee interacting with the explanatory process, given an explanandum and an initial explanans, is called *Explanatory Space* (ES). The explanatory space is defined by:

- p (the explanatory process),
- A (the set of actions),
- D (the explanandum),
- the initial explanans E_0.

Thus, following the formalizations previously presented, the components of an explanatory process are the: Explainee, Instances (X), Ontologies (N), Sources (S).

This is why we say that a good explanans (explanation) has to be bound to both the explainee and the explanandum. In other words, a good ES is:

- Sourced: bound to the sources.
- Adaptable: bound to the specific explainee.
- Grounded: bound to the instances.
- Expandable: bound to the ontologies.

We will call to these properties: the SAGE properties of a good ES.

5.5   Heuristics for Exploring the Explanatory Space

Assuming that the explanandum D is provided as defined above, then:

1. How do we define the explanatory process p?
2. How do we pick the initial explanans E_0?
3. How do we pick the set of actions A?

The answers to these questions are highly dependent to the constraints that the explanatory space is supposed to have. In our scenario, we want to define an explanatory space sufficient to provide user-centred explanations of ADMs for trustworthy AI.

Considering that the narrative explanatory process p consists in exploring an Explanatory Space, in order to define p we identify 3 policies for exploring the ES, namely: the *Simplicity*, the *Abstraction*, and the *Relevance* policies.

Simplicity is mentioned in recommendation 29.5 of the AI-HLEG Policy and Investment Recommendations [12]: "Ensure that the use of AI systems that entail interaction with end users is by default accompanied by procedures to support users [...]. These procedures should be accompanied by simple explanations and a user-friendly procedure". In fact, the amount of information that can be effectively provided to a human is limited by physical constraints. Thus simple explanations about something are more likely to be accepted and understood

and they are better than complicated ones and should be presented earlier than complex ones.

The **Abstraction** policy maintains that preferring abstract concepts over their concretizations helps in keeping the explanation as direct as possible. This is because there might be many concretizations of the same abstract concept. In other words, exploring concretizations before abstractions would generate longer paths in the explanatory space rather than the opposite. If we explore abstractions before concretizations, it is more likely that the ES is less complicated, without losing informative content.

The **Relevance** policy bounds further the explanatory process to the purposes/objectives of the explainee. It states that the information that is more likely to be relevant for the explainee (to reach its objective) should be presented earlier than the less relevant one. One of the expected effects of the relevance policy is that explanations will be shorter.

The Simplicity, Abstraction and Relevance policies shape the explanatory process p, putting significant constraints on the initial explanans E_0 and the set of actions A. Simplicity implies that in the explanatory process we start from a very minimal explanans and iteratively we add information to it in order to make it more detailed and complex. Simplicity also implies that simple representations of the explanandum (e.g. natural language descriptions) should be presented before the original representations (the "Ground"; e.g. XML, JSON, etc..). Abstraction implies that the explanans should start from generic and conceptual information about the explanandum (e.g. its size and other meta-data), going toward more specific and concrete information (the "Ground"). Relevance implies (among other things) that E_0 should contain the most relevant information possible, and that further details should be firstly about the entities directly involved in E_0. The other entities should be explored/presented in a second moment, if needed.

5.6   The Initial Explanans

The initial explanans E_0 should contain an overview of the underlying Explanatory Process (EP), giving explicit information about the purposes (e.g. to give insights about a sequence of events, to verify an hypothesis, etc..) of the EP.[4] E_0 should also provide an overview of the explanandum, pointing to information about the metadata (e.g. size, language, knowledge representation conventions). We will refer to the information contained in E_0 as the "incipit" or "background" of the EP.

In the case of explanations under the GDPR, E_0 should specify (among other things) whether the EP is operating ex-ante or ex-post, and the explanandum referred by E_0 should contain all the information mentioned in Section 2.2. A succinct justification about the automated decision (required in the ex-post

---

[4] It is not excluded that the original purposes might change during the explanatory process.

phase) can be generated through a static explanatory tool such as AIX360 [1], and should be given as part of the EP overview.

In our real-case, the justification about the automated decision states that Giulio's request has been rejected because of the Italian decree.

5.7   The High-Level Actions and the Structure of the Explanatory Space
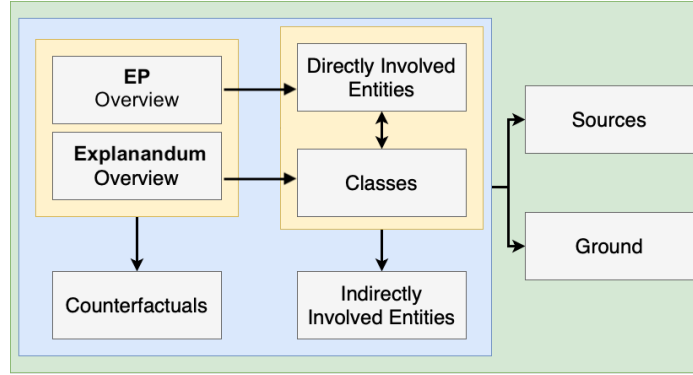

Fig. 2. Structure of the Explanatory Space: arrows show the flow of information, while every rectangle represents a different sub-stage in the Explanatory Space.

The explainee explores the Explanatory Space through a set of pre-defined actions meant to be used to build explanations respecting the SAGE properties. For every SAGE property, we identify a set of commands that may be used by the explainee during the Explanatory Process:

- **Source-ability** (commands): used to show the source of a model (a law, a paper, etc..).
- **Adapt-ability**: used to keep track of important information while exploring the explanatory space, building an argumentation framework.
- **Ground-ability**: used to refer and show specific parts of the explanandum in their original format (e.g. XML, JSON, SQL, etc..), and used for generating counterfactuals.
- **Expand-ability**: used to deepen concepts, aligning them to other concepts available in the explanandum context.

Defining the explanatory process p, the (high-level) set of commands/actions A and the initial explanans E_0, we have also defined the structure of the Explanatory Space. The resulting structure (shown in figure 2) is made of 6 main stages: Incipit (EP and Explanandum Overview), Core Information (Classes and Directly Involved Entities), Marginal Information (Indirectly Involved Entities), Ground, Sources, Counterfactuals. Each stage of the structure is involved in a different step of the explanans construction.

In our real-case the explainee (Giulio) wants to get an explanation of SPIN-dle's conclusions (the process' decision). Giulio uses the Explanatory Process (EP) on the explanandum. The EP starts from the "Incipit" stage (the initial explanans) showing:

- A succinct *ex-post* justification of the process' decision, and that the justification is related to known concepts that can be easily explored.
- The explanandum is made of a knowledge base and a process having a set of inputs that can be changed by Giulio in order to understand whether the justification is valid and useful for him (maybe it is not).

The first level result of EP is too shallow, thus Giulio asks to EP to go down to the "Core Information" stage showing more information about the explanandum. Now Giulio sees that the data is composed by: a set of logical conclusions, the hierarchy of rules used to get those conclusions (the causal chain), the premises on which the rules have been applied. Giulio wants to get more information about the rules, thus it moves to the "Marginal Information" stage, finding out that the Italian decree 101/2018 has rebutted the GDPR and it is responsible for the final decision, thus it marks that information as an argument that supports the process' decision. Furthermore, Giulio sees that every rule is ground-able to a LegalRuleML component, and linkable to the pertinent source of law that justifies the rule. Giulio can also see rebuttals, and if he would ask the EP to tell more about the GDPR rebuttal he would find out that the "Lex specialis derogat generali" is applied, causing the activation of the rule associated to the Italian decree instead of the rule associated to the GDPR.

6 Discussions

With an explanatory tool based on our model, the user can explore the explanatory space and build its own explanatory narrative through a set of predefined actions. The explanatory narrative can be built through the adapt-ability commands, by defining an argumentation framework. The exploration of the explanatory space can be performed through the expand-ability, source-ability and ground-ability commands. The resulting tool is user-centred by design and can be used for finding evidence to make sense of phenomena (sense-making), articulating understandings into an explanatory narrative. Furthermore, we claim that the structure of Explanatory Space we identified is sufficient to produce the descriptive, causal and justificatory explanations required by the GDPR. In fact, assuming that the information at the "Background" stage defines the initial explanans, we have that: Descriptive explanations can be obtained by reasoning over the "Core and Marginal Information" stages of the ES, Causal explanations over the "Counterfactuals" stage mainly, and Justificatory explanations over the "Incipit" and "Sources" stages mainly. It is possible to apply logical rules in order to automatize the production of reasonable explanations. In fact, Description Logic can be used for reasoning over descriptions, Causal Inference

[20] for reasoning over causations, Defeasible Deontic Logic for reasoning over justifications.

Despite this, our model has some limitations. Because it is meant to be effective in building user-centred explanations, but not community-centred ones. In fact, constructing strong, effective and non-over-fitted explanations is historically an iterative and community-centred process.

7   Conclusions and Future Work

We have introduced a model of a User-Centred Narrative Explanatory Process, based on concepts drawn from ISO 9241, as a promising contribution to Trustworthy AI compliant with the GDPR. To this end, we identified a minimal set of required explanation types and we converted the problem of generating explanations into the identification of an appropriate path over an Explanatory Space (ES) defined and eventually generated by every user-centred (interactive) explanatory tool. Finally we provided a reasonable structure of the ES through the identification of the SAGE properties and of some space exploration heuristics. Building a working prototype based upon our model is the next natural step. We are also considering to extend the current model in order to make it suitable also for community-centred explanatory tools.

References

1. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012 (2019)
2. Athan, T., Boley, H., Governatori, G., Palmirani, M., Paschke, A., Wyner, A.Z.: Oasis legalruleml. In: ICAIL. vol. 13, pp. 3–12 (2013)
3. Bennet, W.L., Feldman, M.S.: Reconstructing reality in the courtroom. Tavistock (1981)
4. Berland, L.K., Reiser, B.J.: Making sense of argumentation and explanation. Science Education **93**(1), 26–55 (2009)
5. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., Floridi, L.: Artificial intelligence and the 'good society': the us, eu, and uk approach. Science and engineering ethics **24**(2), 505–528 (2018)
6. Cocarascu, O., Rago, A., Toni, F.: Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1261–1269. International Foundation for Autonomous Agents and Multiagent Systems (2019)
7. Čyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., Greenberg, D., Hapuarachchi, T.: Explanations by arbitrated argumentative dispute. Expert Systems with Applications **127**, 141–156 (2019)
8. Driver, R., Newton, P., Osborne, J.: Establishing the norms of scientific argumentation in classrooms. Science education **84**(3), 287–312 (2000)

9. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al.: Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. Minds and Machines **28**(4), 689–707 (2018)
10. Fox, M., Long, D., Magazzeni, D.: Explainable planning. arXiv preprint arXiv:1709.10256 (2017)
11. HLEG, A.: Ethics guidelines for trustworthy ai (2019)
12. HLEG, A.: Policy and investment recommendations (2019)
13. 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
14. Lipton, P.: What good is an explanation? In: Explanation, pp. 43–59. Springer (2001)
15. Meyer, J.J.C.: Deontic logic: A concise overview. In: Deontic Logic in Computer Science: Normative System Specification, pp. 3–16. Wiley (1993)
16. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018)
17. Norris, S.P., Guilbert, S.M., Smith, M.L., Hakimelahi, S., Phillips, L.M.: A theoretical framework for narrative explanation in science. Science Education **89**(4), 535–563 (2005)
18. Palmirani, M., Governatori, G.: Modelling legal knowledge for gdpr compliance checking. In: JURIX. pp. 101–110 (2018)
19. Passmore, J.: Explanation in everyday life, in science, and in history. History and Theory **2**(2), 105–123 (1962)
20. Pearl, J.: The seven tools of causal inference, with reflections on machine learning. Commun. ACM **62**(3), 54–60 (2019)
21. Prakken, H.: An argumentation-based analysis of the simonshaven case. Topics in cognitive science (2019)
22. Raymond, A., Gunes, H., Prorok, A.: Culture-based explainable human-agent deconfliction. arXiv preprint arXiv:1911.10098 (2019)
23. Sandoval, W.A., Reiser, B.J.: Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. Science Education **88**(3), 345–372 (2004)
24. Suthers, D.D., Toth, E.E., Weiner, A.: An integrated approach to implementing collaborative inquiry in the classroom. In: Proceedings of the 2nd international conference on Computer support for collaborative learning. pp. 275–282. International Society of the Learning Sciences (1997)
25. Verheij, B., Bex, F., Timmer, S.T., Vlek, C.S., Meyer, J.J.C., Renooij, S., Prakken, H.: Arguments, scenarios and probabilities: connections between three normative frameworks for evidential reasoning. Law, Probability and Risk **15**(1), 35–70 (2015)
26. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gpdr. Harv. JL & Tech. **31**, 841 (2017)
27. WP29: Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01). European Commission (2016)
28. Zhong, Q., Fan, X., Luo, X., Toni, F.: An explainable multi-attribute decision model based on argumentation. Expert Systems with Applications **117**, 42–61 (2019)