November 2019

The difference between Explainable and Explaining: requirements and challenges under the GDPR

Francesco Sovrano [a], Fabio Vitali [a] and Monica Palmirani [b]
[a] *DISI, University of Bologna*
[b] *CIRSFID, University of Bologna*

Abstract. We know that Automated Decision-Making (ADM) is currently changing industry, thus people and countries started to be concerned about the impact that may have on everyone lives. The GDPR stresses the importance of a Right to Explanation (e.g., art. 22, artt. 13-14-15, recital 71), requiring the AI industry to adapt consequently, thus giving rise to eXplainable AI (XAI). Modern XAI proposes some solutions to make ADM more transparent following the principle included in the GDPR (art. 5), but many researchers criticize XAI to provide little justification for choosing different explanation types or representations. In this paper we propose a new model of an explanatory process based on the idea of *explanatory narratives*, claiming that it is powerful enough to allow many possible types of explanations including causal, contrastive, justificatory and other types of non-causal explanations.

Keywords. Explanation, Explainability, XAI, HCI

1. Introduction

The academic interest in Artificial Intelligence (AI, [4]) has grown together with the attention of countries and people toward the actual disruptive effects of Automated Decision Making (ADM [20]) in industry and in the public administration (e.g., COMPAS [5]), effects that may affect the lives of billions of people [10]. Thus, GDPR (General Data Protection Regulation, UE 2016/679) stresses the importance of the Right to Explanation, several expert groups, including those acting for the European Commission, requiring the AI industry to quickly adopt ethics code of conducts [2], [7] thus giving rise to eXplainable AI (XAI). So, what is an explanation?

According to many theorists [12], an explanation is "an act intended to make something clear, understandable, or intelligible" [13], while others [9] provide a connection between transparency, traceability, accountability, explanation, justification, interpretation, causation. Additionally the High-Level Expert Group on Artificial Intelligence [14] underlines the fact that the "humans interacting with AI systems must be able to keep full

and effective self-determination over themselves, and be able to partake in the democratic process".
Anyway, the William-Webster Dictionary gives three different meanings to the verb "to explain": i) to make plain or understandable (e.g., "footnotes that explain the terms"), ii) to give the reason for or cause of (e.g., "unable to explain his strange conduct"), iii) to show the logical development or relationships of (e.g., "explained the new theory").
In epistemology the concept of explanations is clearly connected to meaning ii of the William-Webster entry, in that it sees explanations as statements that provide the explainees with understanding of the causes of some facts. This is the meaning that has been adopted quite thoroughly in the field of Explainable Artificial Intelligence (XAI) and exposition of the causes behind the results of the computations have become the prevalent understanding of what explanations need to be [19]. Summarily, according to these points of view, explanations are answers to why-questions, and characterizations and ramifications of why-questions are what are being studied and discussed. Although we see the importance of why-questions in framing the concept of explanation, we find them to be a part of the whole issue, and, in fact, only the *second* part. Before why-questions to provide causal characterization of a fact, we need to consider whether such fact is available and understood for what it actually means. Not only: when dealing with complex systems as in Artificial Intelligence, rarely is a single fact in need of understanding, but, more often, we must deal with a complex web of steps and actions that led the software to come to its conclusion. How to deal with such complexity?

In this paper we discuss a new issue to the problem of explainability and explanation in complex systems and Automated Decision Making systems, including eXplainable Artificial Intelligent Systems (XAI). In fact, differently from much of the current literature on XAI, we do not place any undue weight on the *causal* characterization of explanations, preferring to refer to them as narratives (and, in particular, *explanatory narratives*) generated by an interested narrator/reader that explores a potentially very vast *explanatory space* in order to satisfy a situated and contextualized interest about a complex system.

The rest of the paper is structured as follows. In section 2 we will introduce the background information, pointing to remarkable existing works on explanations, explainability and explaining. In section 3 we introduce an abstract model for an explanatory process that is able to produce many types of causal and non-causal (semantic, contrastive, etc..) explanations. In section 4 we discuss the results, analysing the feasibility of building a real explanatory process based upon our model. Finally, in section 5 we conclude this paper through a brief recap.

2. Background and Related Work

2.1. Explanations as narratives

According to [16], explanations serve to resolve a puzzlement in a listener. And since puzzlements occur in many different ways, explanations serve a number of different functions, ranging from assigning, developing, or expanding meaning, to justifying the facts by appealing to norms, standards, or values, to describing the facts with greater details, to, finally, giving a causal account of these facts. Complexity of facts to be explained also implies that the explanation cannot be a single choice among the above,

but, more often than not, a sequence of explanatory choices shedding light onto the individual facts, but also onto their collection and sequencing in a connected structure [13]. According to [11], explanations also feature *why-regress*, whereby "whatever answer someone gives to a why-question, it is almost always possible sensibly to ask why the explanation itself is so. Thus there is a potential regress of explanations". Explanations therefore shift from simple causal statements of individual events (local dimension) to full explanatory narratives about a complex sequence of facts (global dimension). [3] argues that explanations consider individual instances (*event-tokens*) rather than classes of facts (*event-types*): we explain individual events, not the class this individual event takes part to. Thus "the explanatory pattern [. . .] forms a coherent or connected narrative which represents a number of events[*-tokens*] in an intelligible sequence. Hence the pattern is appropriately called a *narrative explanation*". Narrative explanations naturally must take on features both of narratives and of explanations.

Finally, Norris et al. list eight elements appropriate for framing narrative discourse: "Of primary importance, we believe, are the existence of event-tokens, time, and agency: particular occurrences involving particular actors in the past and over time. The narrator, narrative appetite, structure, purpose, and reader seem to be of secondary importance in determining degree of narrativity". Norris et al. assigns qualifying values to lesser elements, for instance to the nature and goals of the reader, so that: "although these elements are important, we can imagine narratives in which these elements are represented poorly or not at all. [Their absence] signals a story poorly told." Yet, the role of the reader is neither tangential nor optional if we consider, as Lipton does, the issue of familiarity: "explanation is in some sense reduction to the familiar. It is what is strange or surprising that we do not understand; a good explanation gives us understanding by making the phenomenon familiar, presumably by relating it to other things that are already familiar". Thus, an appropriate characterization of the background and goals of the explainees serves to characterize the nature of the explanation sought and to focus and direct the narrative accordingly.

The variety of knowledge backgrounds and of cognitive goals of the explainees, and the need to cater for and oblige to the why-regress that is spontaneous in understanding, mean that the search for the *ideal* or *perfect* explanation is futile and meaningless, that it is the explainees' job to decide whether their goals have been reached, and that subjective satisfaction is the only reasonable metrics to evaluate success in explanation. This makes the quality of explanations as intrinsically relative and wholly dependent on the characterization of the explainees' familiarities and goals. Thus, borrowing the phrase "specified users achieving specified goals [. . .] in a specified context of use" from the definition of usability in ISO 9241-210 2010, we conclusively propose to define explanations as "narratives set to increase understanding over a system or sequence of events for the fulfilment of a specified explainee having specified goals in a specified context of use (e.g., the judge has particular requirements defined by the procedural code: to collect evidences supporting the decision, in the context and point-in-time about the crime, etc.".

2.2. The Right to Explanation

The role of legislations and industry is essential in the quest to increase AI accountability in legal, ethical, policy, and technical terms. Despite this, it seems that these two different streams of work (industry and law) are still working toward different directions, lacking

of common basis to determine the required content of explanations. "As a result, much of the prior work on methods, standards, and other scholarship around explanations will be valuable in an academic or developmental sense, but will fail to actually help the intended beneficiaries of algorithmic accountability: people affected by algorithmic decisions" [17]. As example, the GDPR art. 22 defines the right to claim of a human intervention when a completely automated decision-making system may affect the legal status of a citizen. The 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". This implies that the data subject should access to the whole data, to the logic of the algorithm, to the run-time flow of the processing determining the decision, the context information (e.g., jurisdiction of the ADM) and, in case, also to the datasets used for training, developing and testing (especially in the case of supervised machine learning). It appears that explanations can be offered after decisions have been made (ex-post), and are not a required precondition to contest decisions. It is not completely true: in the artt. 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, as well as the significance and the envisaged consequences of such processing for the data subject." (ex-ante). This combination of those articles make the right of explanation very articulated and composed by different steps. In the *ex-ante* phase we should provide all the information for guaranteeing the transparency principle:

1. The *algorithms* and *models* pipeline composing the ADM.
2. The *data* used for training (if any), developing and testing the ADM.
3. The *context* information (e.g. jurisdiction of the ADM).
4. The possible *consequences* of the ADM on the specific data subject (local principle).

While to contest a decision (during the *ex-post* phase), the data subject should access to:

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

Further, there is no clear link that suggests that explanations under Recital 71 (of the GDPR) require opening the black box but there is a clear indication "to obtain an explanation of the decision reached after such assessment and to challenge the decision".

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. Wachter et al. have assessed three purposes of explanations of automated decisions from the view of the data subject under the GDPR: i) to inform and help the subject *understand* why a particular decision was reached, ii) to provide grounds to *contest* adverse decisions, iii) to understand what could be *changed* to receive a desired result in the future, based on the current decision-making model.

Leveraging on the aforementioned purposes of explanations, counterfactuals are a reasonable way to lawfully provide explanations under the GDPR's right to explanation,

even though counterfactuals are not appropriate to understand system functionality, or the rationale of an automated decision, or to provide the statistical evidence needed to globally assess algorithms for fairness or racial bias.

We believe that explanations should be also contrastive with the explainees' pre-existing knowledge, and they should also be linkable to other domain knowledge, sometimes providing non-causal insights.

3. The Explaining Process and the Properties of a Good Explanatory Space

An explanation is the result of an explaining process that takes place over an input that allows it, as shown in figure 1. Explainable datasets and processes are not the narratives

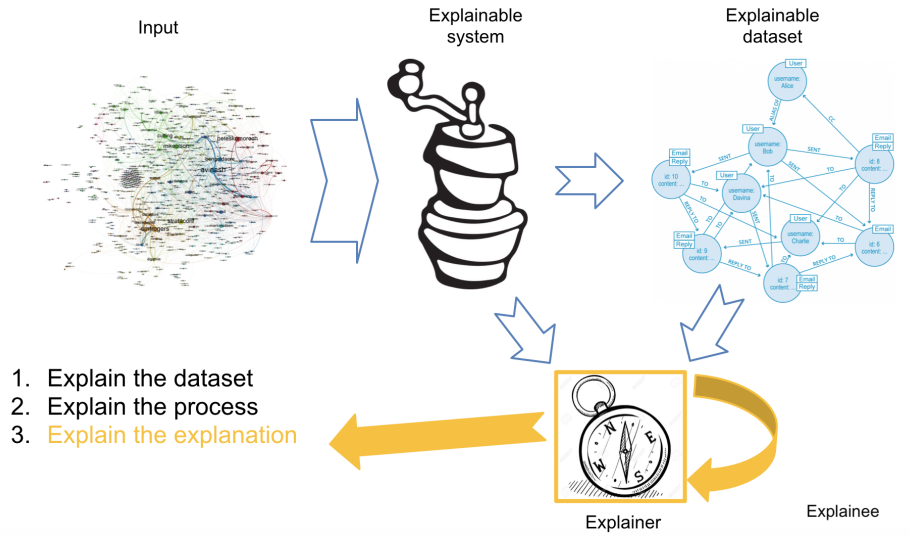

Figure 1.: Explainability vs Explaining

that we seek in explanations, but they require at least "a narrator (someone telling), a narratee (someone receiving [...]), events (something that happened), and past time" [13]. In section 2.1 we introduced our working definition of an explanation as a path in an explanatory space, an explanation as "narratives set to increase understanding over a system or sequence of events for the satisfaction of a specified explainee having specified goals in a specified context of use". The qualities of the explanation that provide the explainee with the necessary satisfaction, following the categories provided by Norris et al. (2005), can be summarized in a good choice of narrative appetite, structure and purpose.

A fully-automated explainer is unlikely to be designed to become the narrator that targets these quality parameters to guarantee the satisfaction of a specified explainee. Rather, we believe that (at least in the case of explanations under the GDPR) readers and narrators must be 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.

Providing the necessary mechanisms for self-told narratives seem therefore a reasonable goal for any explanatory tool, by arranging available information into an *explanatory space* that the narrators-readers can explore according to their whim and goals until the desired level of understanding is reached. According to our model, the input of an explanatory process can be explainable datasets and/or explainable processes (simple and/or higher-order). A dataset is a collection of data items with a type, while a process is a function that takes as input a dataset of parameters/premises, providing as output a dataset of conclusions. Information provided for the explanation can be either global or specific. Global information is related to the concepts of the datasets and the behaviour of the process as a whole, with respect to the totality of the possible values and/or inputs (also known as class-events), while specific information is related to the actual data values and the individual decisions taken by the process considering the specific input instances (also known as token-events). In the case of specific information we may expect that not all the classes of the underlying ontologies, nor all the rules of the algorithm, have been used to get to the final output. The ordered set of rules used to get a local decision is called *causal chain*, and it is plausibly expected to be in the explanatory space in the form of an explainable process.

3.1. Behavioural Model of a good Explanatory Process

To provide a good guidance model for the implementation of a reasonable tool for the exploration of the explanatory space, we introduce here the **SAGES** (Simple, Adaptable, Grounded, Expandable, Sourced) behavioural model of a good explanatory process.

Simple: the explanatory process has to be simple, producing explanations easy to process and understand. This implies that the explanations need to connect with readers with any level of competence, from none whatsoever upwards. In practice this requires creating a fairly ample explanatory space that includes also very basic and fundamental information about the dataset and the process being explained, and some fairly rapid heuristics allowing expert readers to explore the explanatory space in order to find the sought information as fast as possible.

Adaptable: the explanatory process has to be bounded to the appetite and purpose of the narrative. This means that the explainees do not need to be informed on aspects of the dataset and/or process they are not interested in, but can navigate the explanatory space based on the specific goals and context they have.

Grounded: the explanatory process has to be bounded to the specific data and/or process steps involved in the narrative. This implies that the explanatory space must include and give access to the actual data and/or process rules activated during the event, and connect them correctly to the relevant pieces of the narrative. This allows the narrator/reader to build more contrastive explanations.

Expandable: the explanatory process has to be bounded to the classes of the domain. This means that the explanatory space must include and give access to descriptions and explanation elements about the dataset and process domains and any other related domains in general, regardless of the specific event tokens of the narrative. This includes providing access to mechanisms and tools to suggest and explore counterfactuals and their effect on the computation.

Sourced: the explanatory process has to be bounded to provenance specifications of the individual elements of the dataset and the individual process steps, allowing for justificatory explanations of the sequence of events being narrated. Complex rule systems may collect and compose rule-sets of very different origin and authoritativeness, and being able to connect each individual data item and each individual rule to the norms, standards, or values that justify it, is important in order to allow exploration, debate and alternative strategies to put in play with the given XAI system.

3.2. Structure of the Explanatory Space

The SAGES model therefore drives the structure of the explanatory space, organized into seven levels in increasing depth of explanation details. These levels are: Context, Dataset, Classes, Entities, Properties, Grounding, Explorable.

Background level: it provides information describing the characteristics of the context in which the explainer operates. At this level we might find information about the explainer goal (e.g. undefined goal, giving insights about a decision justification, etc..) and whether the explainer is operating ex-ante or ex-post.

Explanandum level: it provides information describing the material characteristics of the input, and its metadata. Among other things, it may contain summary information about: 1. the algorithms and models composing the involved processes; 2. metadata about the physical and virtual context in which the automated process operates; 3. the syntax used in the datasets and the number and the variety of entities described within them.

Classes level: it provides information about the variety of types and classes represented in the dataset or involved in the process.

Entities level: it provides information about each entity described in the dataset with particular attention to human-readable descriptions and fundamental information.

Properties level: it provides information about each property associated to the selected entity as specified in the dataset, and the associated value. At this level we can investigate the provenance of the assertion and of the data, if they are trustworthy.

Grounding level: it explicitly identifies with precision and detail the exact part of the explainable model/dataset associated to the individual narrative element, possibly in the source format in which it was placed in the input to the explaining process. In this part we could explain the logic used (e.g., defeasible logic).

Exploration level: it provides information at the level of the conceptual domains, connecting it with information, facts, norms, values and anything that is not explicitly mentioned in the input but may be relevant for the understanding of the explainee. In this level we could explain the arguments and the counter-arguments.

3.3. Example

We show here a use-case of an explanatory tool based on our model. In this use-case, the 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

[1, 15] using defeasible logic (as shown in figure 2), in order to be able to represent the fact that the GDPR art. 8 rule (16 yearsOld) is overridden with the Italian's one (14 yearsOld). The SPINDle legal reasoner processes the correct rule according to the jurisdiction

```
<lrml:PrescriptiveStatement key="ps3">
  <ruleml:Rule key=":ruletemplate2" closure="universal">
    <lrml:Paraphrase>r3: minor14, italian=> ObtainConsent</lrml:Paraphrase>
    <ruleml:if>
      <ruleml:And key=":and1">
        <ruleml:Atom key=":atom5">
          <ruleml:Rel iri=":minor14"/>
          <ruleml:Var>Y</ruleml:Var>
        </ruleml:Atom>
        <ruleml:Atom key=":atom6">
          <ruleml:Rel iri=":italian"/>
          <ruleml:Var>Y</ruleml:Var>
        </ruleml:Atom>
      </ruleml:And>
    </ruleml:if>
    <ruleml:then>
      <lrml:Obligation iri=":obligation">
        <ruleml:Atom key=":atom7">
          <ruleml:Rel iri=":ObtainConsent"/>
          <ruleml:Var>Y</ruleml:Var>
        </ruleml:Atom>
      </lrml:Obligation>
    </ruleml:then>
  </ruleml:Rule>
</lrml:PrescriptiveStatement>
```

Figure 2.: LegalRuleML of the Art. 8 GDPR.

(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 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 3 shows a possible view of our explanatory tool.

4. Discussions and Future Work

In this paper we proposed a model of an Explainer that takes as input only explainable datasets and/or explainable processes. Our Explainer does not work with uninterpretable processes (e.g. deep neural networks) and datasets. We assumed that this choice is generic enough for our purposes, without providing any strong *scientific* justification, thus without

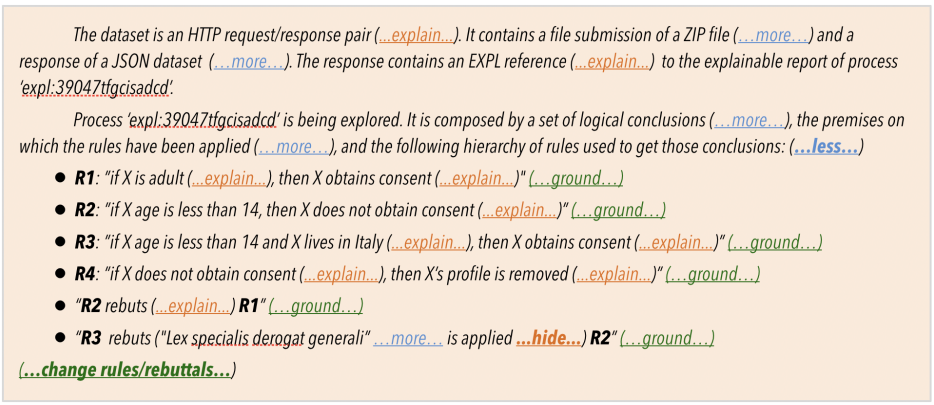

Figure 3.: Explainer of the Art. 8 GDPR application.

properly experimenting the efficacy of our model.
Furthermore it is not clear if our model is generic enough to be compatible with any interface design and any explainable system, and it is not clear how much the model is aligned to the framework for trustworthy artificial intelligence of the High-Level Expert Group on AI (see [6]).
This is why, now, we are going to work toward providing stronger results on the aforementioned issues.

5. Conclusion

Differently from most of the XAI literature, we believe that useful explanations are not only causal explanations. We highlight that many types of explanation may exist, including causal chains, contrastive, justificatory and other types of non-causal explanations. This is why, in this paper, we proposed a new design for the structure of an explanation and the behaviour of and explainer, as an explanatory narrative process. We claim that our explainer is capable of working with both explainable datasets and processes, producing statements that can be rendered to human readers and can be made Simple, Adaptable, Grounded, Expandable and Sourced (SAGES).

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".

References

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

[2] 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.

[3] Carol E Cleland. Methodological and epistemic differences between historical science and experimental science. *Philosophy of Science*, 69(3):447–451, 2002.

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

[5] 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.

[6] Luciano Floridi. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1(6):261, 2019.

[7] 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.

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

[9] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[10] 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.

[11] Peter Lipton. What good is an explanation? In *Explanation*, pages 43–59. Springer, 2001.

[12] 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.

[13] 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.

[14] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. European Commission, 2019.

[15] Monica Palmirani and Guido Governatori. Modelling legal knowledge for gdpr compliance checking. In *JURIX*, pages 101–110, 2018.

[16] John Passmore. Explanation in everyday life, in science, and in history. *History and Theory*, 2(2):105–123, 1962.

[17] Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. *Columbia Business Law Review*, 2019.

[18] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gpdr. *Harv. JL & Tech.*, 31:841, 2017.

[19] 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.

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