Addressing Regulatory Requirements on Explanations for Automated Decisions with Provenance – A Case Study

TRUNG DONG HUYNH, King's College London, United Kingdom
NIKO TSAKALAKIS, University of Southampton, United Kingdom
AYAH HELAL, King's College London, United Kingdom
SOPHIE STALLA-BOURDILLON, University of Southampton, United Kingdom
LUC MOREAU, King's College London, United Kingdom

AI-based automated decisions are increasingly used as part of new services being deployed to the general public. This approach to building services presents significant potential benefits, such as the reduced speed of execution, increased accuracy, lower cost, and ability to adapt to a wide variety of situations. However, equally significant concerns have been raised and are now well documented such as concerns about privacy, fairness, bias and ethics. On the consumer side, more often than not, the users of those services are provided with no or inadequate explanations for decisions that may impact their lives. In this paper, we report the experience of developing a socio-technical approach to constructing explanations for such decisions from their audit trails, or provenance, in an automated manner. The work has been carried out in collaboration with the UK Information Commissioner's Office (ICO). In particular, we have implemented an automated Loan Decision scenario, instrumented its decision pipeline to record provenance, categorized relevant explanations according to their audience and their regulatory purposes, built an explanation-generation prototype, and deployed the whole system in an online demonstrator.

CCS Concepts: • **Theory of computation** → *Data provenance*; • **Security and privacy** → *Information accountability and usage control*; • **Social and professional topics** → *Technology audits*; Automation; Socio-technical systems; **Governmental regulations**.

Additional Key Words and Phrases: explainable computing, GDPR, automated decisions, data provenance

ACM Reference Format:
Trung Dong Huynh, Niko Tsakalakis, Ayah Helal, Sophie Stalla-Bourdillon, and Luc Moreau. 2020. Addressing Regulatory Requirements on Explanations for Automated Decisions with Provenance – A Case Study. *Digit. Gov. Res. Pract.* 2, 1, Article 99 (November 2020), 14 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

AI-based automated decisions are increasingly mediating civic life [21], as they are now routinely used in health, education, justice, employment, finance, the Web and social media, and will soon permeate the functioning of smart cities, governments, and private sector. On the one hand, this approach to building services presents significant potential benefits, such as the reduced speed of execution, increased accuracy, lower cost, and ability

Authors' addresses: Trung Dong Huynh, Department of Informatics, King's College London, Bush House, 30 Aldwych, London, WC2B 4BG, United Kingdom, dong.huynh@kcl.ac.uk; Niko Tsakalakis, University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom, n.tsakalakis@southampton.ac.uk; Ayah Helal, Department of Informatics, King's College London, Bush House, 30 Aldwych, London, WC2B 4BG, United Kingdom, ayah.helal@kcl.ac.uk; Sophie Stalla-Bourdillon, Southampton Law School, University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom, s.stalla-bourdillon@soton.ac.uk; Luc Moreau, Department of Informatics, King's College London, Bush House, 30 Aldwych, London, WC2B 4BG, United Kingdom, luc.moreau@kcl.ac.uk.

2639-0175/2020/11-ART99
https://doi.org/10.1145/1122445.1122456

to adapt to a wide variety of situations. On the other hand, significant concerns have been raised about the risks they present, such as concerns about privacy, fairness, bias and ethics [3, 25]. As a result, several regulatory and legal frameworks have emerged across the world to address some of the concerns arising from automated services. Within the European Union (EU), the data protection framework has been overhauled with the General Data Protection Regulation (GDPR) [7], which includes among others a revised right to information (Articles 13–14), right to access to "meaningful information about the logic involved" in the context of automated decision-making (Article 15), and the right not to be subject to a decision based solely on automated processing (Article 22). Similarly, the move towards strengthening transparency requirements is affecting different parts of the world (cf. the Digital Republic Law[1] in France, the Consumer Privacy Act of 2018[2] in California, and the Modernization of Convention 108[3]).

Not only limited to meeting regulatory obligations, the demand for improved explainability of automated decision-making systems also comes from business and social expectations [22]. For businesses, understanding why a certain decision was made by their systems would help them safeguard their processes against unchecked bias. Failing that may result in significant financial loss and/or reputation damage.[4] For the consumers, explanations about the decisions they received, be it a mortgage loan or a school allocation for their children, would give them confidence in the system. More importantly, such transparency would enable the consumers to contest an automated decision should they believe it was erroneous.

While transparency and accountability are, therefore, starting to be addressed at different levels, a key challenge is that regulatory frameworks remain high-level and do not specify practical means of implementing them, e.g. how the 'logic' of the processing should be derived and expressed [12]. In fact, there is no consensus on what is required in terms of transparency/accountability obligations, on whether transparency necessary leads to fit-for-purpose, *actionable* explanations, or whether/how it is *technically* possible to meet these obligations. To tackle some of those questions, a three-month investigation was initiated by a multi-disciplinary team, consisting of researchers and regulators, formed of computer scientists from King's College London, legal experts from the University of Southampton, and the UK Information Commissioner's Office (ICO). The aim was to seek a concrete approach to help data controllers[5] fulfill some of their obligations under the GDPR concerning explaining aspects of automated decisions to data subjects. This paper reports the socio-technical approach we took to identify and produce GDPR-related explanations for loan decisions and presents the resulted prototype.

Overall, the project demonstrated that explanations for an automated decision under GDPR can be generated from its recorded audit trails, or *provenance* [19]. The provenance of a decision is a form of knowledge graph providing an account of what a system performed to produce that decision — including references to people, data sets, and organizations involved; attribution of data; and data derivations.[6] In the context of automated decision-making, such an audit trail provides valuable information about the individuals, organizations, and factors that influenced the decision, from which explanations on how the system arrived at the decision can be constructed. In this work, we examined an automated decision scenario in which fictitious loan applications are submitted and loan decisions made in an automated manner, similarly to a typical online process of applying for a credit card (see Section 3). Against the scenario, we have:

[1] https://www.republique-numerique.fr/pages/digital-republic-bill-rationale

[2] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375

[3] https://www.coe.int/en/web/data-protection/-/modernisation-of-convention-108

[4] Numerous such incidents have been widely reported in the press; examples include Amazon's facial recognition wrongly identified law makers as criminals (https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28) and Apple Card investigated for gender discrimination (https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html).

[5] Data controllers are those who determines the purposes and means of the processing of personal data in the context of a given application [7, Article 4(7)].

[6] Examples of these are later provided in Figs 2, 3, and 4.

(1) identified thirteen different types of explanations that are meaningful in a GDPR context,
(2) devised an initial set of requirements for application designers to support the automated construction of these explanations,
(3) proposed a technical architecture for generating explanation narratives from provenance, and
(4) built an online demonstrator to produce explanations for loan decisions in the above scenario.

In the remainder of the paper, Section 2 gives an overview of the related work. We then present the loan decision scenario and outline the types of explanations we identified for this scenario (Section 3). Our technical approach for generating explanations from provenance is described in Section 4. We discuss the merits and the limitations of the approach in Section 5, which concludes the paper with some directions for future work.

2 RELATED WORK

When faced with a decision that we find hard to understand, we humans often want an explanation, which could include clarifying information about the process by which the decision was made. This is particularly true in cases where such a decision has an (adverse) impact on us or we believe the decision-making is erroneous. Therefore, as examples, the law in various countries routinely requires judges to explain their rulings and administrative agencies to explain their decisions [5]. Similar obligations also apply to certain private decision-makers in various industries. Consumer reporting agencies in the US, for instance, are required to provide a list of key factors that negatively influenced a consumer's credit score [15].

With the recent widespread adoption of machine learning (ML) in automated and machine-assisted decision-making systems, explanations of their decisions are similarly desired [21]. Recognizing this, in Europe, the GDPR includes a right to access to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing" of personal data [7, Article 15]. Similar provisions are included in the French Digital Republic Act and the Modernization of the Convention 108 (see [6] for a comparison). A key challenge with automated decision-making systems, however, lies in the technical implementation of such obligations given their typical complexity. Exacerbating the problem, some ML models employed by those systems are effectively 'black-box', i.e. there is no apparent explanation for their outputs given the inputs. Hence, it is even more challenging to explain decisions based on the outputs from such models.

The academic community has also recognized the importance of tackling the above explainability concerns, which have become an active research topic world-wide. In the US, the Defense Advanced Research Projects Agency's eXplainable AI program[7] specifically focuses on the explainability of ML-based approaches and quality metrics of these explanations. A number of international events are specifically dedicated to this topic: ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), the Explainable AI workshop at International Joint Conferences on Artificial Intelligence (IJCAI); the Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) workshop.[8] To combat the opaqueness of ML models, a variety of techniques have been recently proposed (see [17] for an overview and [1, 10] for extensive surveys). Approaches include designing the learning process to ensure interpretability of results, approximating a learned model in a more readily intelligible substitution; and offering tools to interact with the model to get a sense for its operation. Notably amongst those, Wachter et al. propose unconditional counter-factual explanations to help understand automated decisions. They do not attempt to clarify how a decision is made internally in a model but instead provide insights into which external factors could be different to arrive at a desired outcome [24]. Instead of using counter-factual cases to explain the causal link between certain variations of a factor and their outcomes, Miller argues for contrastive explanations to show why a decision is chosen in contrast to an alternative by comparing

[7] https://www.darpa.mil/program/explainable-artificial-intelligence.
[8] ACM FAccT conference: https://facctconference.org. FAT/ML workshop: https://www.fatml.org.

their hypothetical outcomes [16]. This approach is adopted in (machine) planning to explain why the planning algorithm chooses a certain plan of actions [13].

All the above techniques, however, focus solely on explaining the behavior of an ML model or an algorithm. Importantly, algorithms and ML models are only one step in the pipelines involved in automated decision-making: data processing and filtering during training an ML model are also critical, as well as dataset selection, which may involve significant human inputs. It is recognized that selecting a dataset of unknown provenance, or dataset with bias, or even poor configuration of the training may result in adverse decisions for people [25]. As a result, the legal context and the scientific community point to the need for broader governance frameworks for automated decisions, including not only explanations of black-boxes but also processes involved in the configuration of such black-boxes, the validation of the results they produce, as well as the ability to demonstrate that due diligence was suitably undertaken. For instance, Rieke et al. propose a framework that highlights how non-technical insights about an automated system (e.g. its designed purpose, the constraining policies that govern human/system behaviors) can be just as important, and often more important, than its technical, tangible artifacts [21]. Likewise, Burt et al. put forward a governance model that applies to automated decision-making pipelines, including its various stakeholders, algorithms, data sets, and surrounding processes [4]. In the same vein, we take a holistic view of a decision-making system and believe that all the above aspects of decision pipelines should be considered when constructing explanations. This is enabled by suitably recording the full audit trail of all the processes that lead to a decision, i.e. its provenance (see Section 4). More recently, the UK ICO published guidance for organizations in the UK on explaining decisions made with AI [23]; it similarly emphasizes holistic principles encompassing organizational, process, and technical aspects when it comes to explaining automated decisions. The provenance-based approach reported in this paper was cited by the ICO as an example approach to producing explanations for the processes involving the data before the black box [23, Section 2, Task 2].

This work has some similarities with the work on generating data narratives by Gil and Garijo, whose aim is to produce an accurate description of scientific workflow executions from their previously recorded provenance [9]. Besides differences in the employed techniques, our approach, however, is to extract requirements for explanations from regulatory obligations and then, based on those, to define requirements on the provenance recording to support them (Section 4). The approach builds on an idea initially developed by Richardson and Moreau of generating natural language from provenance, which focuses on studying linguistics aspects related to meaningful identifiers and their perception by users [20]. Our work, instead, is focusing upon the concept of explainability and its legal groundings. In fact, it can be distinguished from prior research in considering concrete explanations in response to specific legal requirements (from the GDPR in this case), not just generic categories of explanations. Nevertheless, the proposed methodology and technical approach are generalizable and can be extended to applications and laws beyond the scenario studied in this paper.

3 EXPLANATIONS FOR LOAN DECISIONS

Credit applications nowadays are typically assessed by automated systems and often approved or rejected within seconds without human involvement. In order to provide a concrete context for exploring potential questions concerning automated decisions, we created a similar hypothetical loan scenario in which loan applications are decided by an automated pipeline.

3.1 Loan Decision Scenario

Loan Company is a credit institution that offers short-term unsecured loans to borrowers. To minimize loss from *charge-off*, i.e. when a loan is unlikely to be repaid by the borrower, the institution developed a machine-learning pipeline that predicts the probability of charge-off from a loan application. Based on this probability, an automated recommendation is made on whether the application should be approved or rejected. The pipeline was trained

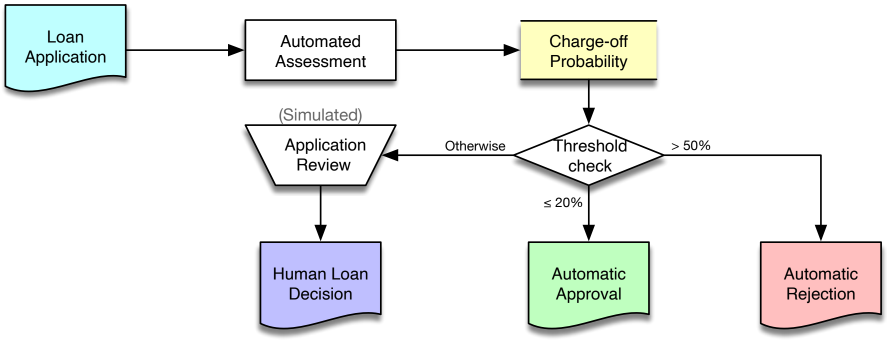


Fig. 1. The flowchart of the simulated loan decision pipeline.

and tested on the company's past loan performance data and was shown to perform reasonably well. It was approved for deployment to assess all future loan applications and is enabled to make automatic decisions in clear-cut cases without the intervention of a loan officer: if the probability of charge-off is higher than 50%, the loan application is automatically rejected; if the probability is less than 20%, it is automatically approved. A loan officer has to examine the remaining cases (i.e. where the probability is between 20% and 50%) and make the final decision. To streamline the demonstrator, such a human decision is simulated in the implemented loan pipeline (depicted in Fig. 1). In GDPR terminology, the Loan Company here is regarded as a *data controller* as it captures personal data from an applicant, regarded as a *data subject*, for a specific purpose (i.e. a loan application).

3.2 Explanation Elicitation Methodology

The above scenario provides us with the necessary details to think concretely about the types of questions one would ask concerning an automated decision produced by such a pipeline. We followed the below steps to elicit various types of explanations relevant to the scenario's stakeholders in two workshops with experts at the ICO:

(1) **Identifying user questions**: Grounded by legal rights afforded to data subjects by the GDPR and other applicable legislation and regulations in the UK (such as the Equality Act 2010[9]), we deliberated a variety of questions that one may ask about the loan decision pipeline and its loan decisions.
(2) **Categorizing questions**: The identified questions were grouped into categories that address the same concerns. For each category, we identified the target audience. Relevant regulatory obligations are also linked with the explanation category, giving the rationale why such an explanation would facilitate meeting those obligations and also, in some cases, business requirements.
(3) **Crafting example answers**: For each category, we brainstormed textual, example answers that address the questions in the category.
(4) **Identifying provenance data requirements**: In each category, using the example answers crafted in Step 3, we identified entities in the universe of discourse that are required to construct such answers. These will serve as the basis for the requirements on provenance data to be recorded to support this explanation category.

Following the above methodology, we identified thirteen concrete categories of explanations, listed in Table 1, each is about a particular aspect of the decision-making. They can be grouped loosely to those that address the concerns of an individual data subject and those that address the concerns of the data controller. Due to the

[9] http://www.legislation.gov.uk/ukpga/2010/15/contents

Table 1. Thirteen categories of explanations identified for decisions in the Loan scenario.

Category	Description	Example Questions
Individual concerns		
Automation	Whether a decision was made without any human involvement.	Has the loan decision been reached solely via automated means?
Data Inclusion	What types of data were used by the pipeline.	What types of data were used to assess my loan application?
Data Exclusion	What types of data were excluded from the decision process.	Which data was ignored and not considered?
Data Source	The origin of a data type.	Where did you get those data about me?
Data Accuracy	Whether the data considered is accurate.	Are the data used for assessing my loan application correct?
Data Currency	How timely relevant the included data.	Is the data used up to date?
Profile-related Fairness	People of a similar profile should be treated similarly in the same process.	Have I been treated similarly to others having the same profile?
Discrimination-related Fairness	Identify any bias against a protected characteristic.	Was I rejected due to my gender?
Institutional concerns		
Performance	Whether the performance of the decision pipeline is satisfactory.	Is the decision pipeline sufficiently accurate?
Responsibility	Who did what and when.	Who decided the data selection?
Process	The processes governing decisions impacting the decision pipeline.	What is the process for choosing the threshold value?
Systemic Discrimination or Bias	Whether the automated decision pipeline exhibit systematic and repeatable unfair treatment to a particular group of data subjects.	Has an equality review carried out on the past loan applications?
Ongoing Monitoring	How often the above were checked and shown to be satisfactory.	When was the pipeline revalidated? How often the accuracy is checked?

limited space, we are not able to include all detailed descriptions of the thirteen categories in this paper; they are available online in our technical report of this work [11]. To give an illustration of an explanation category, we include the description of the Automation category (Table 2) as an example. It addresses the question whether a decision is fully automated without meaningful human involvement. The bold text in parentheses included in the example explanation are identifiers for data and people referred by the explanation; they could be linked to further information accessible in a real application.

4 CONSTRUCTING EXPLANATIONS FOR AUTOMATED DECISIONS

For the purpose of constructing explanations, we assume that audit trails are recorded in systems making automated decisions, enabling us to trace back a decision to its input data and to identify the responsibility for each of the activities found along the way. Such an audit trail is also known as the provenance of the decision. In this work, we adopted the PROV data model (PROV-DM) [19] standardized by the World Wide Web Consortium. Paraphrasing PROV-DM's definition of provenance, we define provenance of a decision as

Table 2. The Automation explanation category.

Audience	Data subjects
Questions	Has the loan decision been reached solely via automated means?
Description	Whether a decision was made solely by automated means without any human involvement.
Rationale	This explanation helps determine whether GDPR Article 22 is applicable and thereby the prohibition applies: "The data subject shall have the right not to be subject to a decision based solely on automated processing..." It is therefore relevant for demonstrating compliance with Article 5(1)(a) (fairness principle) and Article 5(2) (principle of accountability). This explanation should also help understand when best practice as unfolded in Recital 71 is met, e.g. to determine whether either child data or solely automated means have been used. This explanation could also help determine whether the information provided to the data subject as per Article 13, 14 and 15 is adequate.
Examples	No. The automated recommendation was reviewed by a credit officer (`staff/112`) whose decision was based on your application (`applications/34`), the automated recommendation (`recommendation/34`) itself, a credit reference (`credit_history/34`) and a FICO score (`fico_score/34`).

"a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering" that decision. Such a record in the context of automated decisions is a valuable source of data from which to generate explanations about what happened. In this section, we summarize our technical approach for constructing various explanations of automated decisions (Section 4.2), which was developed with the following explanation requirements (**ER**):

ER1 Explanations should be generated from the recorded provenance of an automated decision.
ER2 Explanations must address one or more legal requirement from GDPR.
ER3 Explanations must be computationally tractable.
ER4 Explanations should be understandable by their target audience.

To support them, the provenance of an automated decision must first be recorded with sufficient details as required specifically by the various categories of explanations to be supported (identified by Step 4 in the Explanation Elicitation Methodology, Section 3.2). Generally speaking, the recorded provenance must adhere to the following generic provenance requirements (**PR**); it must allow us:

PR1 to identify the various types of data of the universe of discourse, e.g. loan application, loan applicant, automated or human-based decision, and so on;
PR2 to trace back outcomes to their influencers;
PR3 to attribute or assign responsibility to software systems or humans for actions or outcomes; and
PR4 to identify the various activities, their respective timing, and their contribution to outcomes.

Section 4.1 below demonstrates how the provenance for a decision by the loan decision pipeline (described in Section 3) was modeled to support the above.

4.1 Modeling the provenance of a loan decision

The PROV data model defines three core concepts: *entity*, *activity*, and *agent*; which can be related to one another by PROV relations. In brief, provenance records describe the generation and use of entities (**PR1**) by some activities (**PR4**), which may be influenced in some ways by agents (**PR3**). Some examples of entities in the loan scenario are a loan application, a loan decision, the input data set, the trained ML model, while a loan applicant, a loan officer, the Loan company, for instance, are PROV agents. Activities are actions that happened such as submitting a loan application, training a model, or reviewing an application.

In order to construct explanations about various aspects of an automated decision, we have gathered information about the various influences and processes involved in making the decision. To do so, we instrumented the loan decision pipeline so that the provenance of every step in the pipeline is recorded.[10] Since the full provenance trace of a loan decision is too big to be presented wholly in this paper, we selectively present parts of it to illustrate the provenance modeling.

Fig. 2 shows the three inputs, or entities (depicted by yellow ellipses), used by the pipeline: the loan application (**loan:applications/48**) attributed to the applicant identified by **loan:applicants/48**, his/her credit history (**loan:credit_history/48**) provided by a credit agency, and the FICO score (**loan:fico_score/48**) provided by the FICO organization. Each of the input entities is attributed to the responsible agent (depicted by orange pentagons) via an attribution relation.

In the loan pipeline, the inputs were transformed and combined into a set of features of the loan application (identified by **py:loan_features/48** in Fig. 3), which was then used in the activity **ex:classify_loans/48** (depicted by a blue rectangle) to generate the automated recommendation **ex:recommendation/48**. The (automated) activity was carried out by a computer (**ex:machine/98. . .**) on behalf of the Loan Company and it was using a pre-trained ML pipeline (**loan:pipeline/1**).[11] In this particular case, the probability of charge-off is higher than the automatic approval threshold, the application was referred to a loan officer to a review. The provenance of the review process is provided in Fig. 4, where it shows that the final loan decision was attributed to an officer (**loan:staff/112**) whose review activity used all the input data entities in addition to the automated recommendation to reach that decision.

The provenance of a loan decision recorded in our demonstrator is hence a knowledge graph that allows one to trace from the decision back to the input data (**PR2**) and to identify the responsibility (**PR3**) for each of the activities (**PR4**). For the sake of brevity, the full provenance graph of the decision (provided in the Supplementary Materials) is split into three simplified graphs (Figs. 2, 3, and 4); it is, in fact, a single, connected graph (**PR2**). In the graph, each of the entities, activities, and agents in the provenance is annotated by types using one or more **prov:type** attributes. Most types are application-specific such as **ln:LoanApplication**, **ln:FICOScore**, and **ln:CreditReference**. In addition, we tagged certain entities with types that will allow for identification of relevant data in support of explanation generation: **pl:Controlled**, **pl:HumanLedActivity**, **prov:SoftwareAgent**, **prov:Person**, and so on (**PR1**). These are highlighted by red dotted boxes in Figs. 2, 3, and 4.

4.2 Generating explanations from provenance

In this section, we present the technical approach for constructing explanations from the provenance of a decision providing that it is recorded in a suitable manner (with respect to the provenance requirements) The approach is summarized in Fig. 5. The full provenance record, albeit necessary for auditing and other purposes, itself is not conducive to construct an explanation directly; it typically contains too many details that a user may find irrelevant, tedious, or overwhelming. Instead, the provenance needs to be processed to produce relevant information nuggets in support of a specific explanation's purpose. For each explanation category (see Table 1), we define an *explanation template* in support of a specific question about an automated decision; it consists of two

[10] In systems where logs and audit trails are sufficiently recorded, provenance information can be instead generated with the provenance template approach [18] without the need for instrumentation.

[11] We also recorded the full provenance of the loan decision pipeline, which is provided in the Supplementary Materials.

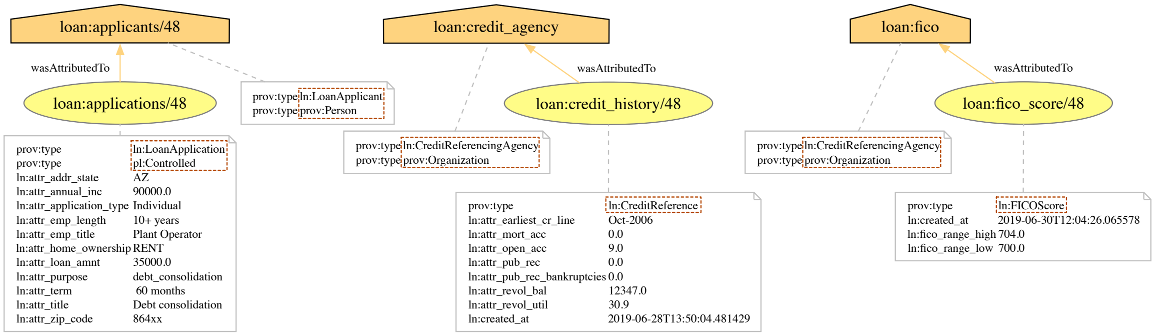

loan:applicants/48
wasAttributedTo
loan:applications/48
prov:type ln:LoanApplicant
prov:type prov:Person
prov:type ln:LoanApplication
prov:type pl:Controlled
ln:attr_addr_state AZ
ln:attr_annual_inc 90000.0
ln:attr_application_type Individual
ln:attr_emp_length 10+ years
ln:attr_emp_title Plant Operator
ln:attr_home_ownership RENT
ln:attr_loan_amnt 35000.0
ln:attr_purpose debt_consolidation
ln:attr_term 60 months
ln:attr_title Debt consolidation
ln:attr_zip_code 864xx
loan:credit_agency
wasAttributedTo
loan:credit_history/48
prov:type ln:CreditReferencingAgency
prov:type prov:Organization
prov:type ln:CreditReference
ln:attr_earliest_cr_line Oct-2006
ln:attr_mort_acc 0.0
ln:attr_open_acc 9.0
ln:attr_pub_rec 0.0
ln:attr_pub_rec_bankruptcies 0.0
ln:attr_revol_bal 12347.0
ln:attr_revol_util 30.9
ln:created_at 2019-06-28T13:50:04.481429
loan:fico
wasAttributedTo
loan:fico_score/48
prov:type ln:CreditReferencingAgency
prov:type prov:Organization
prov:type ln:FICOScore
ln:created_at 2019-06-30T12:04:26.065578
ln:fico_range_high 704.0
ln:fico_range_low 700.0

Fig. 2. Provenance describing some input data and their origins (to be used by the activity in Fig. 3).

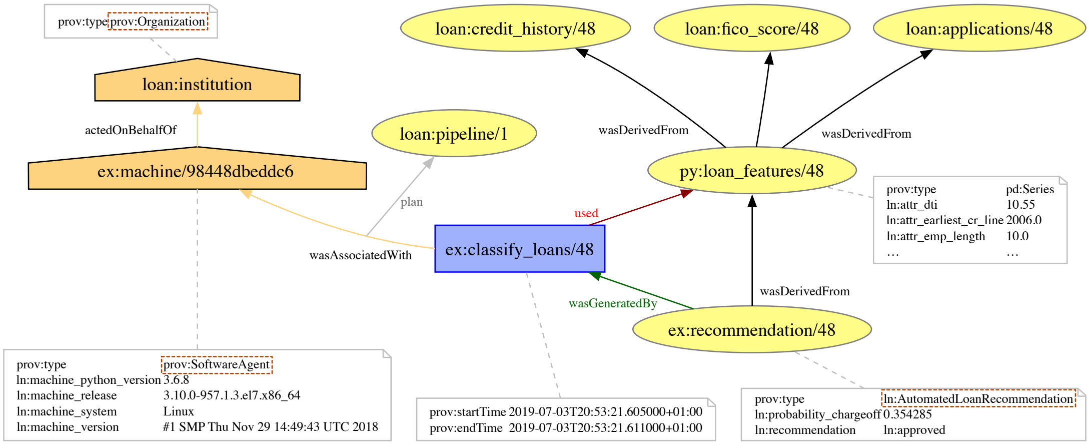

prov:type prov:Organization
loan:institution
actedOnBehalfOf
ex:machine/98448dbeddc6
loan:pipeline/1
plan
wasAssociatedWith
loan:credit_history/48
loan:fico_score/48
loan:applications/48
wasDerivedFrom
wasDerivedFrom
py:loan_features/48
used
ex:classify_loans/48
prov:type pd:Series
ln:attr_dti 10.55
ln:attr_earliest_cr_line 2006.0
ln:attr_emp_length 10.0
… …
wasGeneratedBy
wasDerivedFrom
ex:recommendation/48
prov:type prov:SoftwareAgent
ln:machine_python_version 3.6.8
ln:machine_release 3.10.0-957.1.3.el7.x86_64
ln:machine_system Linux
ln:machine_version #1 SMP Thu Nov 29 14:49:43 UTC 2018
prov:startTime 2019-07-03T20:53:21.605000+01:00
prov:endTime 2019-07-03T20:53:21.611000+01:00
prov:type ln:AutomatedLoanRecommendation
ln:probability_chargeoff 0.354285
ln:recommendation ln:approved

Fig. 3. Provenance of an automated recommendation (`ex:recommendation/48`) by the loan pipeline.

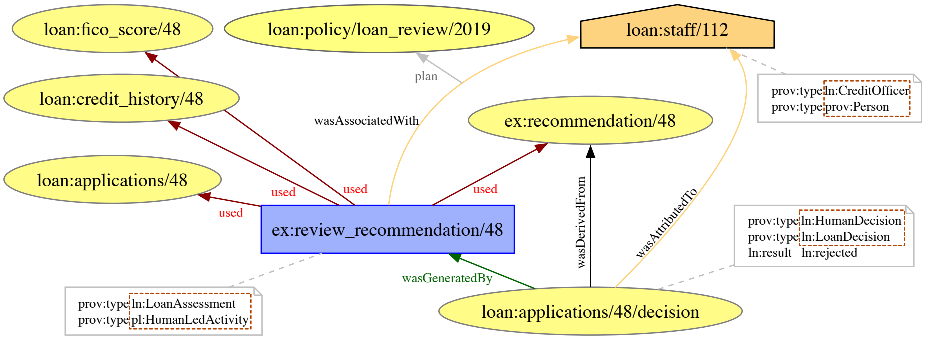

loan:fico_score/48
loan:policy/loan_review/2019
loan:staff/112
plan
loan:credit_history/48
prov:type ln:CreditOfficer
prov:type prov:Person
wasAssociatedWith
ex:recommendation/48
loan:applications/48
used
used
used
used
ex:review_recommendation/48
wasDerivedFrom
wasAttributedTo
prov:type ln:HumanDecision
prov:type ln:LoanDecision
ln:result ln:rejected
wasGeneratedBy
prov:type ln:LoanAssessment
prov:type pl:HumanLedActivity
loan:applications/48/decision

Fig. 4. Provenance of a loan decision (`loan:application/48/decision`) made by a credit officer (`loan:staff/112`) based on the input data (shown in Fig. 2) and the automated recommendation (shown in Fig. 3).

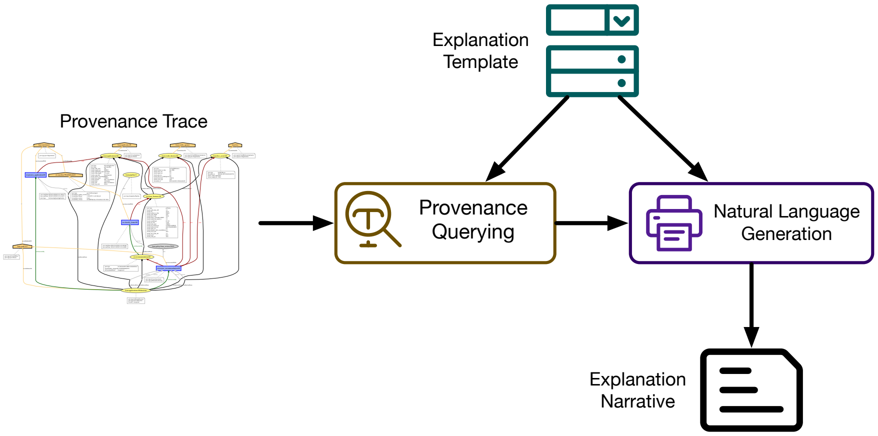


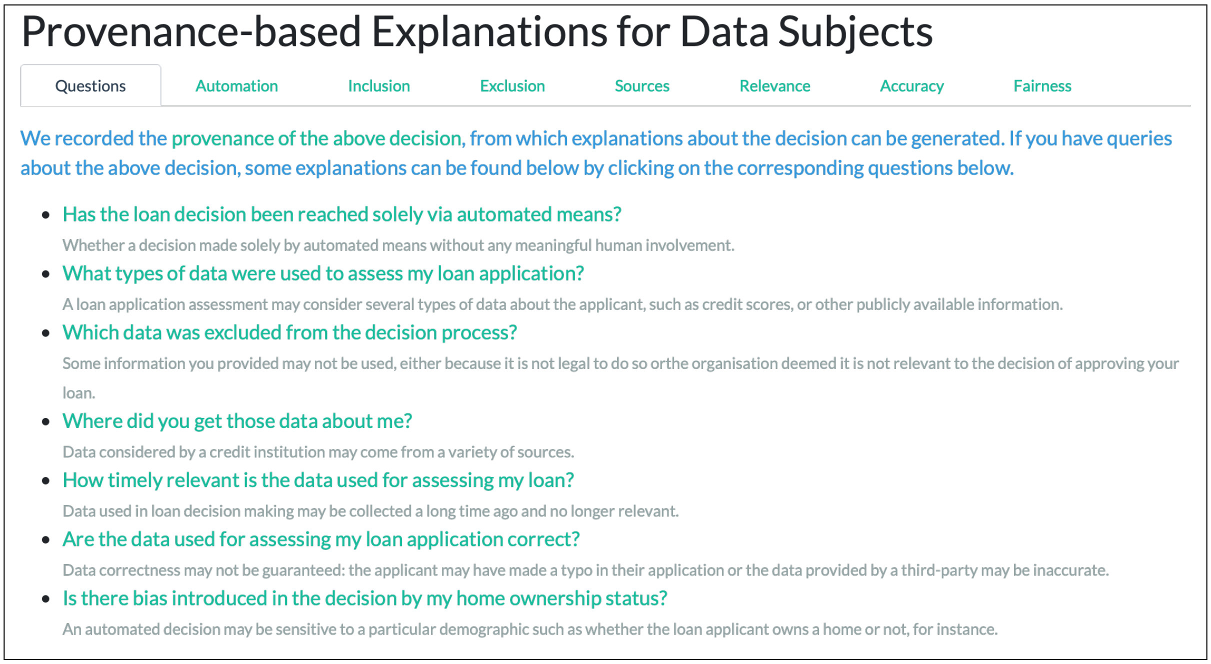
Fig. 5. Generating explanation narratives from provenance — an overview.

parts: a provenance query and a narrative template. The former determines which elements of the provenance graphs are relevant to the chosen explanation and refers to specific annotated types (**PR1**, as highlighted in Figs. 2, 3, and 4) to help find them in the provenance trace. The latter is a natural language text template inspired from the example answers identified in Step 3 of the Explanation Elicitation Methodology (Section 3.2); it contains placeholders to be filled with information extracted by the provenance query. Hence, the generation of an explanation narrative from the provenance (of a decision) involves two steps:

(1) Given an explanation template, the query part of the template is executed over the decision's provenance trace to extract specific parts of the full graph into a smaller provenance graph to be used in the next step.
(2) Information contained in the extracted sub-graph is used to complete the corresponding narrative template; the result is then processed by a natural language generation (NLG) engine[12] to construct the sentences constituting the explanation narrative.

More details of the above are provided in [11], the technical report of this work. The whole approach was implemented and packaged with the provenance-instrumented loan decision pipeline into a technical demonstrator, available online at explain.openprovenance.org/loan. The website allows users to play the role of a customer applying for a loan, who would go through filling in an application form, submitting the application, and finally receiving a decision. The decision is presented along with several questions that the customer can pose about various aspects of the decision (see Fig. 6). These are the questions we have identified for each explanation category (Step 1 of the Explanation Elicitation Methodology, Section 3.2). For instance, they may ask whether the decision was solely automated; the Automation tab provides the user with the answer (shown in Fig. 7), which is generated in real-time from the recorded provenance of the decision. Below each explanation, the legal rationale that grounds the explanation is also provided. Note that the demonstrator's interface, as shown in Figs. 6 and 7, was mainly designed to make a range of explanations and rationales available; it was not aimed to be representative of a real loan application website.

[12] We use the SimpleNLG library [8] as the NLG engine in the demonstrator.



Provenance-based Explanations for Data Subjects

We recorded the provenance of the above decision, from which explanations about the decision can be generated. If you have queries about the above decision, some explanations can be found below by clicking on the corresponding questions below.

- Has the loan decision been reached solely via automated means?
  Whether a decision made solely by automated means without any meaningful human involvement.
- What types of data were used to assess my loan application?
  A loan application assessment may consider several types of data about the applicant, such as credit scores, or other publicly available information.
- Which data was excluded from the decision process?
  Some information you provided may not be used, either because it is not legal to do so orthe organisation deemed it is not relevant to the decision of approving your loan.
- Where did you get those data about me?
  Data considered by a credit institution may come from a variety of sources.
- How timely relevant is the data used for assessing my loan?
  Data used in loan decision making may be collected a long time ago and no longer relevant.
- Are the data used for assessing my loan application correct?
  Data correctness may not be guaranteed: the applicant may have made a typo in their application or the data provided by a third-party may be inaccurate.
- Is there bias introduced in the decision by my home ownership status?
  An automated decision may be sensitive to a particular demographic such as whether the loan applicant owns a home or not, for instance.

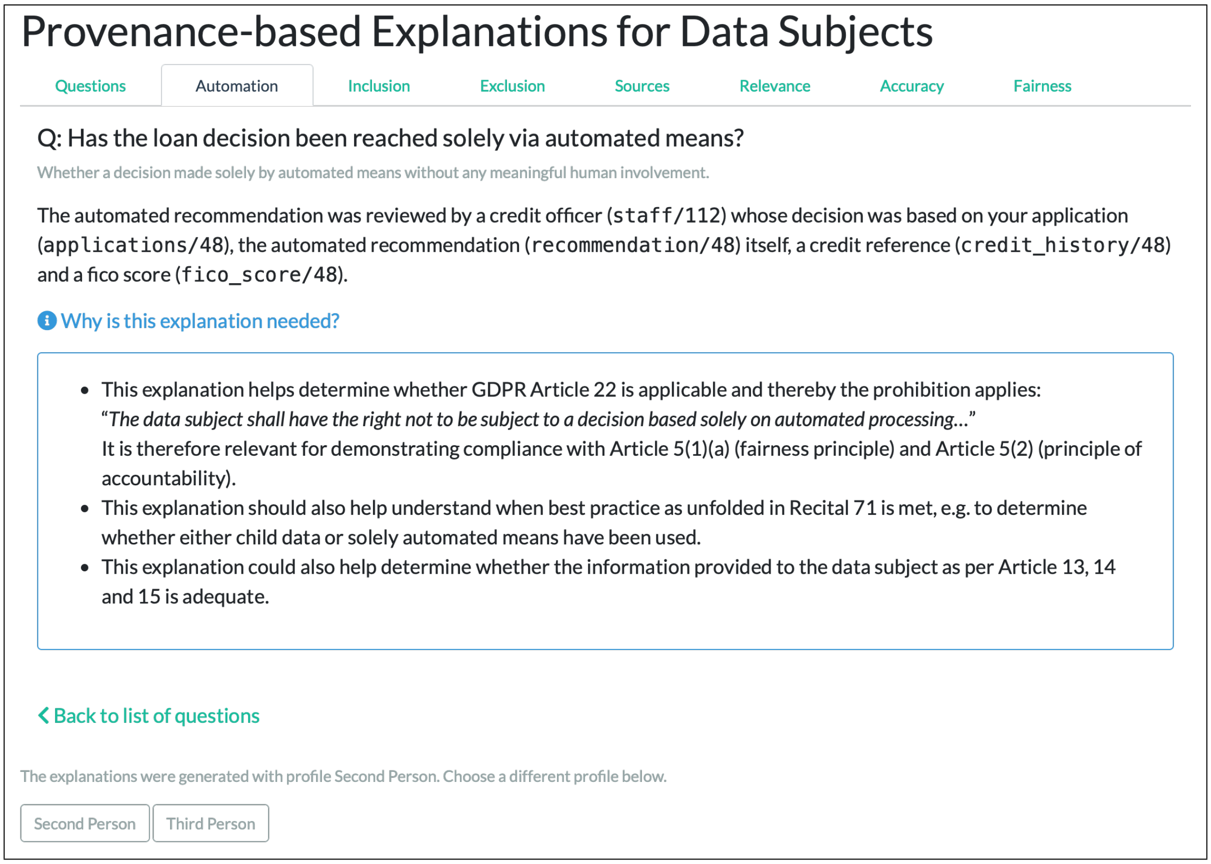
Fig. 6. The questions from the explanation categories are offered to a customer when a loan decision is returned.

5 DISCUSSION & CONCLUSION

Via the explanations for loan decisions presented by the demonstrator, overall, we have demonstrated that provenance of a decision — the knowledge graph capturing the influences, data dependencies, and processes underpinning the decision — provides a solid foundation for generating its explanations. Out of the thirteen categories of explanations, the demonstrator supports eight, all of which are generated from provenance information (**ER1**) in real-time (**ER3**); each of them originated from expert analysis of the legal requirements from GDPR (**ER2**). We have not explored the fairness and bias explanation categories due to lack of consensus on their concrete definitions. We could also have supported the Process and Ongoing Monitoring explanations but they require provenance data that cannot be recorded within the demonstrator and need to be documented from (human) processes outside the pipeline.

This work was carried out within a short period of time (three months) and it has some limitations:

- Our focus was on the legal analysis and technical feasibility; we left the evaluation of the acceptance of the constructed explanations by their target audience (**ER4**) to the future work. Moreover, explanations can and should be refined to fully meet their purposes. Suitable requirement capturing and user studies will help validate these.
- We designed the above prototype for one application scenario, for one ML pipeline, for one specific regulatory framework (GDPR), and for a subset of its requirements. We intend to generalize the approach to other scenarios, regulations and requirements in our future work.
- The approach is predicated on finding certain mark-ups in the provenance to be able to construct the relevant explanations. Besides the above generalization, there is also a clear need to document such mark-ups, so that organizations can adapt their systems to produce suitably annotated provenance. It has to



Provenance-based Explanations for Data Subjects

Q: Has the loan decision been reached solely via automated means?

Whether a decision made solely by automated means without any meaningful human involvement.

The automated recommendation was reviewed by a credit officer (`staff/112`) whose decision was based on your application (`applications/48`), the automated recommendation (`recommendation/48`) itself, a credit reference (`credit_history/48`) and a fico score (`fico_score/48`).

ℹ Why is this explanation needed?

- This explanation helps determine whether GDPR Article 22 is applicable and thereby the prohibition applies:
  "*The data subject shall have the right not to be subject to a decision based solely on automated processing...*"
  It is therefore relevant for demonstrating compliance with Article 5(1)(a) (fairness principle) and Article 5(2) (principle of accountability).
- This explanation should also help understand when best practice as unfolded in Recital 71 is met, e.g. to determine whether either child data or solely automated means have been used.
- This explanation could also help determine whether the information provided to the data subject as per Article 13, 14 and 15 is adequate.

< Back to list of questions

The explanations were generated with profile Second Person. Choose a different profile below.

Fig. 7. The Automation explanation produced by the online demonstrator.

be understood by them that a failure to generate provenance with the right mark-ups will result in the system's inability to construct some explanations.

- Some aspects of the decision-making pipeline are currently not explained. It is particularly the case of the ML algorithm itself, which remains a black-box: the algorithm was used to create a model and the model was used to classify loan applications. Both the model creation and classification are modeled by activities in the recorded provenance. If some libraries can generate further, more detailed provenance for those activities, this, in turn, can be used to construct explanations for them.

To address some of the above concerns, we have started a follow-on research project: Provenance-driven and Legally-grounded Explanations for Automated Decisions (PLEAD). It will extend the initial investigation in this work to three different scenarios: credit rating, school admission allocation, and re-use of data obtained under investigatory powers. The project will study a variety of regulatory frameworks relevant to these scenarios and carry out user engagements with stakeholders.[13]

In addition, this work has also opened up a number of interesting research questions that require further investigation.

[13] More information about the PLEAD project is available at its website: plead-project.org.

- **Automation**: We generate different explanations for automated and human decisions based on information recorded in their provenance. However, it is not clear how we can determine whether any human involvement in a decision is *meaningful*. How much is added by the human on top of the automated recommendation they proceed? Can meaningfulness be determined in an automated manner? Which semantic mark-up in the provenance would help with this task?
- **Data exclusion**: We were able to demonstrate that some loan application characteristics (or elements of third-party data such as credit reference) were not used by the decision-making pipeline. This information, while certainly useful, is looking at "syntactic usage": certain data may have been passed to the pipeline but may or may not have been effectively used to reach a decision. In other words, the data may or may not have influenced the final decision. Such information can only be surfaced if we gain a better understanding of the black box, for example, by calculating the marginal contributions of the input data [14].
- **Forms of explanations**: We present explanations in this work as text sentences answering specific questions. What are other media or forms that can be utilized to effectively deliver and present the information we extract from the provenance of a decision? For instance, explanations, potentially, could be given as part of a dialogue between the system and its targeted recipients [2] or in a graphical representation (e.g. a Gantt chart, a data flow chart). Which form/medium is best suited to which category of explanations?

This work is only the start of a journey, with many directions and research questions lying ahead. While technology underpinning automated decision-making is a source of concerns, we believe that technology also has a place to help address them. The solution should not solely be addressed by technology, but instead, technology must certainly be part of the solution, particularly because compliance should ideally be performed speedily, with accuracy, and at the lowest cost possible. With that in mind, this work has shown that provenance information provides the technological foundations to generate explanations for an automated decision and, by so doing, makes the processes that surround a "black-box" model more transparent and accountable. When provenance of an automated decision is suitably recorded, it becomes possible to computationally query the provenance graph and extract the relevant information to construct the desired explanation for that decision. Those include explanations about the processes that led to the decision being made, who was responsible for what step in these processes, whether the ML model was solely responsible for the decision, what data from which source influenced the decision, and so on. Not only such explanations would help organizations demonstrate compliance to their regulatory obligations, but they would also help improve the confidence of their users in their business processes. Finally, the approach is applicable not only for ML pipelines but also for any form of computing activity requiring explanations where provenance can be recorded.

ACKNOWLEDGMENTS

The work presented in this paper has been supported by the UK Engineering and Physical Sciences Research Council (EPSRC) via the Impact Acceleration Account at King's College London (Grant EP/R511559/1) and via the EPSRC Grant EP/S027238/1 for the PLEAD project.

We would like to thank the ICO's Strategic Policy Projects team for hosting two workshops, sharing their expertise, and providing valuable inputs throughout this work.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Floris Bex and Douglas Walton. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation* 7, 1 (jul 2016), 55–68. https://doi.org/10.3233/AAC-160001

[3] Reuben Binns. 2017. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research: Conference on Fairness, Accountability and Transparency* 81 (dec 2017), 149–159. arXiv:1712.03586

[4] Andrew Burt, Stuart Shirrell, Brenda Leong, and Xiangnong George Wang. 2018. *Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models*. Technical Report. Future of Privacy Forum. https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf

[5] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, Adrian Weller, and Alexandra Wood. 2019. *Accountability of AI Under the Law: The Role of Explanation*. arXiv:1711.01134

[6] Lilian Edwards and Michael Veale. 2018. Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy* 16, 3 (may 2018), 46–54. https://doi.org/10.1109/MSP.2018.2701152 arXiv:1803.07540

[7] European Union. 2016. Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* 59, L 199 (2016), 1–88. http://data.europa.eu/eli/reg/2016/679/oj

[8] Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation* (Athens, Greece) *(ENLG '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 90–93.

[9] Yolanda Gil and Daniel Garijo. 2017. Towards Automating Data Narratives. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI '17*. ACM Press, New York, New York, USA, 565–576. https://doi.org/10.1145/3025171.3025193

[10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (jan 2019), 1–42. https://doi.org/10.1145/3236009 arXiv:1802.01933

[11] Trung Dong Huynh, Sophie Stalla-Bourdillon, and Luc Moreau. 2019. *Provenance-based Explanations for Automated Decisions: Final IAA Project Report*. Technical Report. King's College London. 27 pages. https://kclpure.kcl.ac.uk/portal/en/publications/provenancebased-explanations-for-automated-decisions(5b1426ce-d253-49fa-8390-4bb3abe65f54).html

[12] Margot E. Kaminski. 2019. The Right to Explanation, Explained. *Berkeley Technology Law Journal* 34, 1 (2019), 189–218. https://doi.org/10.15779/Z38TD9N83H

[13] Benjamin Krarup, Michael Cashmore, Daniele Magazzeni, and Tim Miller. 2019. Towards Model-Based Contrastive Explanations for Explainable Planning. In *2nd ICAPS Workshop on Explainable Planning (XAIP-2019)*. 21–29.

[14] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[15] Michael F. McEneney and Karl F. Kaufmann. 2005. Implementing the FACT Act: Self-Executing Provisions. *The Business Lawyer* 60, 2 (2005), 737–747.

[16] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (feb 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007 arXiv:1706.07269

[17] Christoph Molnar. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

[18] Luc Moreau, Belfrit Victor Batlajery, Trung Dong Huynh, Danius Michaelides, and Heather Packer. 2018. A Templating System to Generate Provenance. *IEEE Transactions on Software Engineering* 44, 2 (feb 2018), 103–121. https://doi.org/10.1109/TSE.2017.2659745

[19] Luc Moreau and Paolo Missier. 2013. *PROV-DM: The PROV Data Model*. Technical Report. World Wide Web Consortium. http://www.w3.org/TR/2013/REC-prov-dm-20130430/ W3C Recommendation.

[20] Darren P. Richardson and Luc Moreau. 2016. Towards the Domain Agnostic Generation of Natural Language Explanations from Provenance Graphs for Casual Users. In *Provenance and Annotation of Data and Processes. IPAW 2016*, Marta Mattoso and Boris Glavic (Eds.). Lecture Notes in Computer Science, Vol. 9672. Springer, Cham, 95–106. https://doi.org/10.1007/978-3-319-40593-3_8

[21] Aaron Rieke, Miranda Bogen, and David G. Robinson. 2018. *Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods*. Technical Report. Upturn and Omidyar Network. 40 pages. https://omidyar.com/public-scrutiny-of-automated-decisions-early-lessons-and-emerging-methods/

[22] The Royal Society. 2019. *Explainable AI: The Basics*. 32 pages. https://royalsociety.org/topics-policy/projects/explainable-ai/

[23] The UK Information Commissioner's Office. 2020. *Explaining decisions made with AI*. Technical Report. 136 pages. https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/

[24] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2017). https://doi.org/10.2139/ssrn.3063289

[25] Tal Zarsky. 2016. The Trouble with Algorithmic Decisions. *Science, Technology, & Human Values* 41, 1 (jan 2016), 118–132. https://doi.org/10.1177/0162243915605575