Explainable AI: From Arguments to Decisions

Barbara Futyma, Eric Nakoja, David Pomerenke, Laurens Rutten, David Schimmel, JingYang Zeng
Supervised by: Dr. ir. ing. Nico Roos and Dr. Pieter Collins
Department of Data Science and Knowledge Engineering, Maastricht University

***Abstract*—We propose to make decision support systems more explainable by augmenting them with argumentation systems to provide a human-understandable explanation of the decisions. The core of our project will be a prototype to create an explainable decision tree from an input set of defeasible rules. The result could be used in public and private legal services, as well as medical applications.**

I. Motivation

Artificial Intelligence (AI) decision support systems and their cutting-edge technologies have proven their value to the world. Research in medical diagnostic systems, image and voice recognition, autonomous vehicles, etc. have all adopted AI to achieve optimum performance. However, there is a well-known aspect of Artificial Intelligence that is lacking: The explanation of the final output.

Legal decisions are based on vast and complex bodies of rules. Decision support systems are a valuable help in improving these decisions. With the increasing usage of legal technology, the relevance of the explainability of the decisions made by these systems is becoming clear, and a research community around explainable artificial intelligence has emerged.

Decisions made by public organs in many states need to be explainable due to the requirement of the rule of law. Judges in the United Kingdom, France, and Germany are required to explain their decisions, and decisions may be rescinded on the sole basis of a lack of explanation. In the United States, this requirement has traditionally been more relaxed; it has been tightened recently in both the US and France. Not only judges but also administrative agencies are bound in many cases to explain their decisions, especially if they affect the rights of individual people.[1] This, in turn, calls for the explainability of automated reasoning. In the Netherlands, the *Rechtbank Den Haag* has recently ruled that an "[a]n administrative organ that partially bases its actions on [a system that a human user does not understand] is unable to properly justify its actions and to substantiate its decisions."[2] Moreover, the highest Dutch administrative court, *Raad van State* has recently decided that in order to apply automated decision systems for governmental decisions, they need to be fully comprehensible for judges as well as for normal citizens.[3]

Companies, in contrast, in principle need not justify their actions. The General Data Protection Regulation (GDPR) in the EU has recently introduced a *right to explanation* even towards companies [4]: In certain cases, especially related to personal profiling as used for the calculation of credit scores, clients are entitled to "obtain an explanation of the decision reached after such assessment and to challenge the decision" [5]. This has been criticized as insufficient, and there is an ongoing debate about more advanced regulation concerning explainability.[6] The market for legal technology is substantial: While we are not aware of quantification of the global market, the investment in the innovation of legal technology is estimated at 260 million GBP in the United Kingdom in 2019.[7]

Another important area where decisions can have a critical impact on the lives of humans is the field of medicine. Unexpected diagnostic results need human acceptance before administering to patients due to legal implications and standards in operations. For instance, Craven et al. [8] presented an assumption-based argumentation system and applied it in a medical domain where medical knowledge was derived from the results of clinical trials of early-stage breast cancer treatments and was represented by defeasible rules. They investigated if their system can find arguments whether a daily course of a specific drug for two years was justified for an example patient.

The main reasons in favor of explainability are increasing trust and detecting and fixing possible errors in the algorithms.[1] Explainability is commonly held to consist of transparency (comprising simulatability, decomposability, and algorithmic transparency) and interpretability (comprising textual description, visualization, local explanations, and examples).[9] Analyses of various machine learning methods yield that the most suitable methods with regard to explainability are decision trees[10][9] and approaches based on deductive logic[9].

A novel approach towards explainable artificial intelligence is to augment decision support systems, which are typically black box systems, with argumentation systems. The principle idea behind this is that argumentation systems allow for the extraction of the reasoning leading to its conclusion. If the decision system can be mapped to the argumentation system in such a way that the reasons for the decision correspond to the propositions of the arguments, an explanation behind a decision can be extracted. To formally account for the possibility of exceptions from rules in decision making, the logic underlying the argumentation system can be extended by defeasible rules.

A common implementation of argumentation systems relies on the tableaux proof method. In that vein, Fox et al. [11] propose augmenting decision support systems with defeasible argumentation logic utilizing the ASPIC-framework as a framework for implementing schemas of how certain arguments can support certain actions or plans as a result of

the decision support system. We want to expand on that idea and provide a concrete implementation of a decision support system augmented by an argumentation system utilizing defeasible logic in semantic tableaux proofs. However, we do not want to rely on an external framework like ASPIC.

II. Concepts

A. Defeasible logic

Defeasible logic is a non-monotonic logic, which means that the conclusions drawn are tentative and can be invalidated when new information is introduced. In defeasible logic, defeasible rules are used. Those rules can be defeated by other rules. When a rule is defeated then the conclusion drawn from it is no longer valid. There are multiple versions of defeasible logic. We stick to the definition of Roos [12]: The preference relation defines the priority among the rules thus, the more preferred rule defeats the one with lower priority.

B. Theorem proving

Truth tables are unsuitable for theorem proving due to their strictly exponential complexity. There are four common methods for theorem proving: Sequent calculus[13], calculus of natural deduction [13], [14] (both used mostly by philosophers and other humans), resolution [15] (used mostly by computers), and semantic tableaux [13], [16] (used by both humans and computers).[17]

Semantic tableaux rely on a specific ruleset for the expansion of nodes, and this flexibility makes them applicable to many different kinds of logic. Rulesets have been devised for propositional logic, predicate logic, modal logic, dynamic logic,[17], doxastic logic,[17], epistemic logic,[17] and natural logic[18], among others. Recently, a version for defeasible logic has been proposed[16]. The expectation is that the tree-like structure of a semantic tableaux makes extracting the relevant reasons for why an argument holds or is refuted straight-forward.

C. Argumentation systems

Argumentation systems are a popular tool for the analysis of complex arguments in artificial intelligence in the legal domain. An argumentation system consists of inference rules, arguments, argument structures, and completeness conditions. The argument structure is the set that contains relating arguments. These arguments are connected by the inference rules in the form of a tree. The completeness conditions indicate the boundaries of the arguments.

An *abstract argumentation system* is defined as a set that contains some arguments and the expressed relations among the arguments.[19] Although researchers have developed various abstract argumentation systems,[20] we start by considering abstract argumentation as proposed by Dung[21] for its generality. For our project, such an argumentation system would need to be innovated in terms of explainability and uncertainty.

Many Abstract argumentation systems emphasize accuracy and efficiency, however, interpretability is significant as well[22], because the comprehensive abstracting process and argument results can improve the credibility of decision.

III. State of the art

A. ASPIC+

One framework for structured argumentation that is widely used is ASPIC+ [23]. ASPIC+ allows its users to generate abstract argumentation frameworks in the sense of Dung [21]. It allows the user to freely choose the logical language, inference rules (strict and defeasible), knowledge bases (axioms and ordinary premises), and the argument preference ordering. When it is given these parameters it can tell the user how arguments can be built within the rules, how they can be attacked, and how these attacks can be resolved. This is very useful for this project since it allows the user to see how the system comes to its conclusion.

The main limitation of ASPIC+ is that it is not a system but a framework used for specifying systems. The user still has to define the rules of the system and the reasoner, ASPIC+ only gives insight into the conclusions.

B. Reasoner implementations

The most prominent use of *resolution* is in the Prolog programming language. However, Prolog is technically limited in that it only supports Horn clauses rather than a full logic, and in that it uses depth-first search.[15] While there are extensions to meet both problems, the more fundamental problem in the context of this project is its lack of providing explanations. A Prolog-like logic programming language based on semantic tableaux called *Tablog* has been proposed, but no implementation could be found.[24] Moreover, resolution-based theorem provers for predicate logic have been developed in Python.[1][2]

Notable for *defeasible logic* is the *Oscar* system developed by Pollock, which has the high aim of creating a general rational agent.[25] It is implemented in LISP,[3] and uses natural deduction for propositonal and predicate logic, with an extension for defeasible logic.[26], [27] It is based on a variant of defeasible logic pioneered by Pollock, relying on the concepts or undercutting and rebutting defeaters.[25] This is incompatible with our project, since we use a defeasible logic based on a priority relation, as defined by Roos.[12]

To our knowledge, there are no relevant publications regarding the implementation of *semantic tableaux*, but there are several open-source projects and libraries. For propositional logic, documented implementations are available in Haskell[4], Scala[5][6], and Javascript[7]. For predicate logic, there is a Prolog

[1] https://github.com/yakuza8/first-order-predicate-logic-theorem-prover
[2] https://github.com/evhub/pyprover
[3] https://johnpollock.us/ftp/OSCAR-web-page/oscar.html
[4] https://github.com/ghulette/semantic-tableaux
[5] https://gist.github.com/anrizal06/1239945
[6] http://voidmainargs.blogspot.com/2011/09/semantic-tableaux-in-less-than-90-lines.html
[7] https://github.com/DACUS1995/Semantic-Tableaux-Method-for-Propositional-Logic

paper[28], a Haskell package[8] and an implementation in Javascript [9] (cf. [29]).

IV. Problem statement & research questions

This project aims to investigate the translation of formal specifications from the legal domain (and possibly other domains) into automated decision systems.

Based on this problem statement, we formulate the following research questions:

- Is it practically feasible to augment a decision support system with an argumentation system based on defeasible logic?
- What are the requirements for the explanations that must be provided by an automated decision system?
- Is it possible to provide an adequate explanation with each decision?
- Can legal specifications, such as the tax laws, be used to automatically generate a decision system that provides explanations with the decisions made?
- How well can the system be applied to other fields? How does the degree to which the rules informing the argumentation system are formalized influence its applicability?
- What adaptations are required when applying the system to other fields?
- How detrimental are problems regarding the formal interpretation of rule sets for the reliability and explainability of the decisions?

V. Approach & implementation

The approach proposed here is to follow up on the idea of making explanations behind decisions concluded by automated decision support systems available by augmenting the decision support system with an argumentation system. Rather than relying on available frameworks for argumentation systems, we propose to directly integrated solution for bringing the two systems together. As a basis for the argumentation system, we propose utilizing the semantic tableaux proof method for defeasible logic. The expectation is that the tree-like structure of a semantic tableaux makes extracting the relevant reasons for why an argument holds or is refuted straight-forward. And defeasible logic is expected to alleviate the problem of multiple incompatible states that are all non the less admissible with regards to the argument. The conditions of successfully carrying over the insights provided by the argumentation system to the decision support system will, then, be the primary focus of the research project.

VI. Deliverables & requirements

To evaluate the progress in investigating the research questions, we suggest the following milestones and deliverables. A schematic overview of these items and the time allocated to them is provided in Fig. 1:

- The first major deliverable is the development of an application **prototype** in Python that can evaluate arguments via semantic tableaux method and keep track of the reasoning behind the corresponding judgments. It features a decision support system that is based on this argumentation system and provides the reasoning for each possible set of propositions supporting the decision. The minimum goal set for this project is to have this prototype developed and working for defeasible predicate logic.[10] Extending the application to also support defeasible predicate logic is aimed at, yet depends on the overall progress during the project time. The goal is to have a rudimentary prototype by the end of phase 2 and extend the functionality as far as possible during the third project phase. Thus, the explainable model of abstracting argument is one of our research goals.
- This prototype is accompanied by the second major deliverable: A **research report** summing up the insights gained during the research project. Related to the core functionality provided by the prototype, we strive to investigate further into approaches for formalizing argumentation systems and for how they can be used to supplement decision support systems with the aims of improving the robustness of our prototype, and being able to more clearly compare our approach to alternatives.
- A third deliverable is a small or large **dataset of formalized laws** (and, possibly, rules from other areas) that can be used as an input to our system, and to provide a benchmark for similar future contributions. Possibly, we will just reuse an existing dataset; possibly, we will not find much data and we will restrict ourselves to toy examples. In any case, we will aim to make use of the data from Cremer's thesis [30], adapting it for both defeasible propositional and defeasible predicate logic.

These major deliverables are supplemented by a set of minor deliverables:

- Since the evaluation of the functionality is important for delivering a well-tested prototype and can also motivate possible areas of application, an optional goal is to **investigate ways of retrieving arguments** from real-world contexts like juridical or medical texts. The underlying idea is that formally regulated contexts provide easier means of also formally representing these ideas. We aim to include in our report a survey of existing approaches as well as a glance over alternatives. This rounds out the situation of our prototype in the most likely area of application. If during the research a particularly available approach is found, we aim to implement that at least in parts, with the added benefit of not only showing the viability of the prototype, but also its behaviour within these contexts.
- To round out the development of the prototype and prepare it for handing it over to the interested stakeholder,

[8] http://hackage.haskell.org/package/tableaux,
online demo: http://kashmir.dcc.fc.up.pt/cgi/tableaux.cgi

[9] https://github.com/wo/tpg

[10] We have hosted our current version of this prototype at https://xai.davidpomerenke.vercel.app/

we reserve time to formulate a **developer documentation** and a **user documentation**.

The developer documentation consists of code commentary that gets incrementally added during development, as well as a detailed description of the dependencies and set-up of the prototype. This will be handed in together with the project report.

Part of the user documentation is a layman's website, giving an easy-to-understand overview of the project. The first version of this website will be provided towards the end of period two. It will be updated upon achieving each major step. If desired, a more detailed manual describing the usage can be provided.

- Also related to rounding out the product is to set up the development repository as an **open source** project. This includes organizing the different products as Github repositories, which will also be used during the development. If there is sufficient time at the end, we will also aim to create a programming library and publish it at a Python package repository such as *PyPI*.
- Since the overarching goal of explainable artificial intelligence is to empower the end-user who is not acquainted with programming, it would make sense to create a web **user interface** where the user can enter arguments and facts and obtain the respective decision tree in an intuitive way. This is also an optional milestone since, from the perspective of the artificial intelligence student and the research community, this step is negligible.

VII. Risk analysis & contingency plan

It is unclear whether the implementation of the basic prototype implementing defeasible predicate logic is straightforward enough to leave enough time for implementing it. Because of that uncertainty, the basic prototype will be developed from the ground up with possible extensions like predicate logic in mind. That means ensuring the extensibility by suitable means like properly encapsulating functionalities and giving clear indications of how to extend the program.

The inclusion of test data depends on the availability and accessibility of the test data itself, or of easy approaches for their generation. A precursory investigation leaves doubt whether an adequate data set can be found. If that turns out to be problematic, we will manually include various toy examples that are used or inspired by common examples found in the literature of argumentation systems.

In a similar vein, giving a practical evaluation of how useful the results produced by the prototype are, depends on both the progress of development of the prototype itself, as well as the availability of data sets from different areas. If this practical evaluation appears inadequate to achieve with the products we obtain, then at least an informed literature-based evaluation should be given.

References

[1] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, A. Weller, and A. Wood, "Accountability of AI Under the Law: The Role of Explanation," *arXiv:1711.01134 [cs, stat]*, Dec. 2019, arXiv: 1711.01134. [Online]. Available: http://arxiv.org/abs/1711.01134 (visited on 09/27/2020).

[2] Rechtbank den Haag, *ECLI:NL:RBDHA:2020:1878 (English)*, nl, Last Modified: 2020-03-09, Mar. 2020. [Online]. Available: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 (visited on 10/02/2020).

[3] Raad van State, *ECLI:NL:RVS:2017:1259*, nl, Last Modified: 2020-03-03, May 2017. [Online]. Available: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RVS:2017:1259 (visited on 09/09/2020).

[4] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"," en, *AI Magazine*, vol. 38, no. 3, pp. 50–57, Oct. 2017, Number: 3, ISSN: 2371-9621. DOI: 10.1609/aimag.v38i3.2741. [Online]. Available: https://www.aaai.org/ojs/index.php/aimagazine/article/view/2741 (visited on 09/27/2020).

[5] N. Vollmer, *Recital 71 EU General Data Protection Regulation (EU-GDPR)*, en, text, Publisher: SecureDataService, May 2020. [Online]. Available: http://www.privacy-regulation.eu/en/recital-71-GDPR.htm (visited on 09/27/2020).

[6] L. Edwards and M. Veale, "Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?" *IEEE Security Privacy*, vol. 16, no. 3, pp. 46–54, May 2018, Conference Name: IEEE Security Privacy, ISSN: 1558-4046. DOI: 10.1109/MSP.2018.2701152.

[7] Thomson Reuters and Legal Geek, *Legaltech Startup Report 2019—A Maturing Market*, en-GB, Section: Technology & Innovation, Oct. 2019. [Online]. Available: https://blogs.thomsonreuters.com/legal-uk/2019/10/18/a-new-report-legaltech-startup-report-2019-a-maturing-market/ (visited on 10/02/2020).

[8] R. Craven, F. Toni, C. Cadar, A. Hadad, and M. Williams, "Efficient argumentation for medical decision-making," in *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[9] B. Waltl and R. Vogl, "Explainable artificial intelligence the new frontier in legal informatics," *Jusletter IT*, vol. 4, pp. 1–10, 2018.

[10] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," en, *Information Fusion*, vol. 58,

pp. 82–115, Jun. 2020, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253519308103 (visited on 09/27/2020).

[11] J. Fox, D. Glasspool, D. Grecu, S. Modgil, M. South, and V. Patkar, "Argumentation-Based Inference and Decision Making–A Medical Perspective," *IEEE Intelligent Systems*, vol. 22, pp. 34–41, Nov. 2007. DOI: 10.1109/MIS.2007.102.

[12] N. Roos, "On Resolving Conflicts Between Arguments," *Computational Intelligence*, vol. 16, no. 3, pp. 469–497, 2000, ISSN: 1467-8640. DOI: 10.1111/0824-7935.00120. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/0824-7935.00120 (visited on 10/06/2020).

[13] Open Logic Project, *The Open Logic Text*, Complete build, Revision 2451c5. [Online]. Available: https://github.com/OpenLogicProject/OpenLogic/tree/52451c5cfb55d621e395e0009253f4f6f505902a.

[14] P. D. Magnus, *forall x. An Introduction to Formal Logic*, en, Version 1.4. [Online]. Available: https://github.com/OpenLogicProject/forallx/tree/fa620bb41d9fbb33dc2437878f0072a7c4dcbf51.

[15] S. J. Russell, P. Norvig, and E. Davis, *Artificial intelligence: a modern approach*, en, 3rd ed, ser. Prentice Hall series in artificial intelligence. Upper Saddle River: Prentice Hall, 2010, 37211, ISBN: 978-0-13-604259-4.

[16] N. Roos, "A Semantic Tableau Method for Argument Construction," *BeNeLux AI Conference (BNAIC)*, 2020.

[17] ——, "Logics for Artificial Intelligence," Lecture script, Maastricht University, Oct. 2019.

[18] L. Abzianidze, "LangPro: Natural Language Theorem Prover," *arXiv:1708.09417 [cs]*, Aug. 2017, arXiv: 1708.09417 version: 1. [Online]. Available: http://arxiv.org/abs/1708.09417 (visited on 09/26/2020).

[19] G. A. W. Vreeswijk, "Abstract argumentation systems," en, *Artificial Intelligence*, vol. 90, no. 1, pp. 225–279, Feb. 1997, ISSN: 0004-3702. DOI: 10.1016/S0004-3702(96)00041-0. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370296000410 (visited on 10/02/2020).

[20] H. Prakken, "An abstract framework for argumentation with structured arguments," en, *Argument & Computation*, vol. 1, no. 2, pp. 93–124, Jun. 2010, ISSN: 1946-2166, 1946-2174. DOI: 10.1080/19462160903564592. [Online]. Available: http://content.iospress.com/doi/10.1080/19462160903564592 (visited on 09/10/2020).

[21] P. M. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games," en, *Artificial Intelligence*, vol. 77, no. 2, pp. 321–357, Sep. 1995, ISSN: 00043702. DOI: 10.1016/0004-3702(94)00041-X. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/000437029400041X (visited on 09/26/2020).

[22] Z. Zeng, C. Miao, C. Leung, and C. J. Jih, "Building More Explainable Artificial Intelligence with Argumentation," en, p. 2,

[23] S. Modgil and H. Prakken, "The ASPIC + framework for structured argumentation: A tutorial," en, *Argument & Computation*, vol. 5, no. 1, pp. 31–62, Jan. 2014, Publisher: IOS Press, ISSN: 1946-2166. DOI: 10.1080/19462166.2013.869766. [Online]. Available: https://content.iospress.com/articles/argument-and-computation/869766 (visited on 09/10/2020).

[24] Y. Malachi, Z. Manna, and R. Waldinger, "TABLOG: The deductive-tableau programming language," in *Proceedings of the 1984 ACM Symposium on LISP and functional programming*, ser. LFP '84, New York, NY, USA: Association for Computing Machinery, Aug. 1984, pp. 323–330, ISBN: 978-0-89791-142-9. DOI: 10.1145/800055.802049. [Online]. Available: https://doi.org/10.1145/800055.802049 (visited on 10/02/2020).

[25] J. L. Pollock, "OSCAR: A general-purpose defeasible reasoner," en, *Journal of Applied Non-Classical Logics*, vol. 6, no. 1, pp. 89–113, Jan. 1996, ISSN: 1166-3081, 1958-5780. DOI: 10.1080/11663081.1996.10510868. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/11663081.1996.10510868 (visited on 10/02/2020).

[26] ——, "IV. Sentential reasoning," in *Oscar Manual*. [Online]. Available: https://johnpollock.us/ftp/OSCAR-web-page/MANUAL/MANUAL-Chapter-4.pdf (visited on 10/02/2020).

[27] ——, "V. First-order reasoning," in *Oscar Manual*. [Online]. Available: https://johnpollock.us/ftp/OSCAR-web-page/MANUAL/MANUAL-Chapter-5.pdf (visited on 10/02/2020).

[28] J. Posegga and P. H. Schmitt, "Implementing semantic tableaux," in *Handbook of Tableau Methods*, Springer, 1999, pp. 581–629.

[29] Ş. Minică, "RAESON: A Tool for Reasoning Tasks Driven by Interactive Visualization of Logical Structure," *arXiv:1507.03677 [cs]*, Jul. 2015, arXiv: 1507.03677. [Online]. Available: http://arxiv.org/abs/1507.03677 (visited on 09/26/2020).

[30] T. Cremers, "Defeasible Logic as Professional Support for Regulation Analysis and Creation and Validation of the Specification of the Corresponding Services.,"

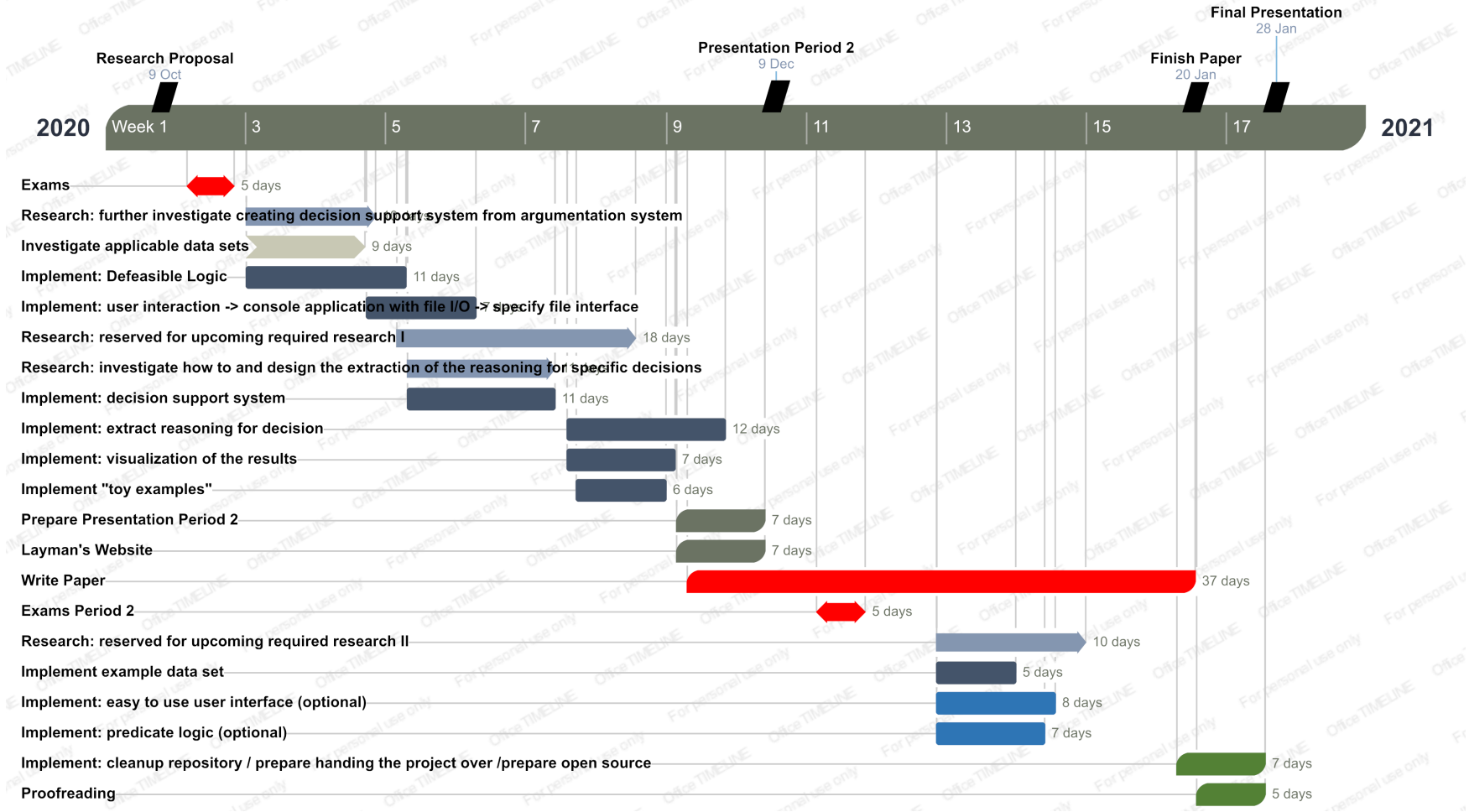

Fig. 1: Milestones and deliverables.