Interactive Natural Language Technology for Explainable Artificial Intelligence

Jose M. Alonso[1](✉), Senén Barro[1], Alberto Bugarín[1], Kees van Deemter[2], Claire Gardent[3], Albert Gatt[4], Ehud Reiter[5], Carles Sierra[6], Mariët Theune[7], Nava Tintarev[8], Hitoshi Yano[9], and Katarzyna Budzynska[10]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
{josemaria.alonso.moral,senen.barro,alberto.bugarin.diz}@usc.es

[2] Universiteit Utrecht, Utrecht, Netherlands
c.j.vandeemter@uu.nl

[3] Lorraine Research Laboratory in Computer Science and its Applications (LORIA), Centre National de la Recherche Scientifique (CNRS), Nancy, France
claire.gardent@loria.fr

[4] Institute of Linguistics and Language Technology, University of Malta (UM), Msida, Malta
albert.gatt@um.edu.mt

[5] Department of Computing Science, University of Aberdeen (UNIABDN), Aberdeen, Scotland
e.reiter@abdn.ac.uk

[6] Artificial Intelligence Research Institute (IIIA), Spanish National Research Council (CSIC), Bellaterra, Spain
sierra@iiia.csic.es

[7] Human Media Interaction, Universiteit Twente (UTWENTE), Enschede, Netherlands
m.theune@utwente.nl

[8] Department of Software Technology, Technische Universiteit Delft (TU Delft), Delft, Netherlands
n.tintarev@tudelft.nl

[9] MINSAIT, INDRA, Madrid, Spain
hyano@minsait.com

[10] Laboratory of The New Ethos, Warsaw University of Technology, Warsaw, Poland
Katarzyna.Budzynska@pw.edu.pl

Abstract. We have defined an interdisciplinary program for training a new generation of researchers who will be ready to leverage the use of Artificial Intelligence (AI)-based models and techniques even by non-expert users. The final goal is to make AI self-explaining and thus contribute to translating knowledge into products and services for economic and social benefit, with the support of Explainable AI systems. Moreover, our focus is on the automatic generation of interactive explanations

Supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621.

F. Heintz et al. (Eds.): TAILOR 2020, LNAI 12641, pp. 63–70, 2021.
https://doi.org/10.1007/978-3-030-73959-1_5

in natural language, the preferred modality among humans, with visualization as a complementary modality.

Keywords: Explainable Artificial Intelligence · Trustworthiness · Multi-modal explanations · Argumentative conversational agents · Human-centered modeling · Human-machine persuasive interaction

1 Introduction

According to Polanyi's paradox [25], humans know more than they can explain, mainly due to the huge amount of implicit knowledge they unconsciously acquire through culture, heritage, etc. The same applies to Artificial Intelligence (AI)-based systems, which are increasingly learnt from data. However, as EU laws also specify, humans have a right to explanation of decisions affecting them, no matter who (or what AI-based system) makes such decisions [23].

In addition, it is worth noting that the European Commission identified AI as the most strategic technology of the 21st century [10] and eXplainable AI (XAI for short) has fast become a fruitful research area. In particular, CLAIRE[1], a Confederation of Laboratories for AI Research in Europe, emphasizes "the need of building trustworthy AI that is beneficial to people through fairness, transparency and explainability". In addition, TAILOR[2], one of four AI networks in the H2020 program ICT-48 "Towards a vibrant European network of AI excellence centres", has the purpose of developing the scientific foundations for Trustworthy AI in Europe. Moreover, Explainable Human-centred AI is highlighted as one of the five key research areas to consider and it is present in five out of the eight pilot experiments to be developed in the H2020 AI4EU project[3] that is funded by call ICT-26 2018.

However, even though XAI systems are likely to make their impact felt in the near future, there is a lack of experts to develop fundamentals of XAI, i.e., researchers ready to develop and to maintain the new generation of AI-based systems that are expected to surround us soon. This is mainly due to the inherently multidisciplinary character of this field of research, with XAI researchers and practitioners coming from very heterogeneous fields. Moreover, it is hard to find XAI experts with a holistic view as well as sufficient breadth and depth in all related topics. Therefore, the H2020-MSCA-ITN project "Interactive Natural Language Technology for Explainable Artificial Intelligence" (NL4XAI[4]) is developing an outstanding training program, deployed across scientific lead institutions and industry partners, with the aim of training 11 creative, entrepreneurial and innovative Early-Stage Researchers (ESRs) who are learning how to develop trustworthy self-explanatory XAI systems. NL4XAI is

[1] https://claire-ai.org/.
[2] https://liu.se/en/research/tailor.
[3] https://www.ai4eu.eu/.
[4] https://nl4xai.eu/.

the first European training network with a focus on Natural Language (NL) and XAI. In the NL4XAI program, ESRs are trained in the fundamentals of AI, along with Computational Linguistics, Argumentation and Human-Machine Interaction Technologies for the generation of interactive explanations in NL as a complement to visualization tools.

In this position paper, we describe how the NL4XAI project contributes to strengthening European innovation capacity in the XAI research area. The rest of the manuscript is organized as follows. Section 2 introduces the research objectives in the NL4XAI project. Section 3 sketches the training program for ESRs. Finally, Sect. 4 provides readers with final remarks.

2 Research Challenges for Early-Stage Researchers

The NL4XAI training program covers four main research objectives:

– **Designing and developing human-centred XAI models**. This objective is addressed by four ESRs who face the challenges of (1) explaining current AI algorithms (paying special attention to the explanation of decision trees and fuzzy rule-based systems [2–5], logical formulas [20], counterfactual facts [32,33], knowledge representation and reasoning, temporal and causal relations in Bayesian networks [18,24,29], and black-box machine learning algorithms [15,19] such as deep neural networks [17]); and (2) generating new self-explanatory AI algorithms. Explanations in NL, adapted to user background and preferences, will be communicated mainly through multi-modal (e.g., graphical and textual modalities) via interactive interfaces to be designed as a result of achieving the next three objectives.
– **Enhancing NL Technologies for XAI**. This objective refers to the need to go deeper with the generation of explanations in NL, as humans naturally do. Two ESRs focus on the study of NL technologies, regarding both NL processing (NLP) and NL generation (NLG) [16,22]; but paying special attention to the grounding of symbolic representations in multi-modal information, and to text production and verbalization in both data-driven/neural and knowledge-based systems [11]. Such systems can be end-to-end or modular [12,21], in the latter case incorporating well-understood NLG sub-tasks such as content determination, lexicalization, linguistic realization, etc. A promising format for NL explanations is in the form of a narrative or story, which is known to aid human comprehension [13,36].
– **Exploiting Argumentation Technology for XAI**. This objective deals with analyzing advantages and drawbacks of current argumentation technology in the context of XAI [6,26,30]. Two ESRs address the challenge of how to organize naturally the discourse history in either narrative/story or dialogue, in terms of standard and customized argumentation schemes [35]; with the focus on designing argumentation-based multi-agent recommender systems [34] that are expected to be self-explainable, non-biased and trustworthy [7].

– **Developing Interactive Interfaces for XAI**. This objective refers to the communication layer between XAI systems and humans [1]. Three ESRs research multi-modal interfaces [31] (i.e., with the goal to generate explanations based on textual and non-verbal information such as graphics/diagrams, but also gestures by embodied conversational agents) associated to virtual assistants in different application domains (e.g., e-Health); with an emphasis on how to convey non-biased, effective and persuasive explanations through verbal and non-verbal interaction between XAI systems and humans [28].

3 Training Program

Each ESR has a principal supervisor in his/her host institution, a secondary supervisor in another institution of the NL4XAI consortium, and a secondment supervisor (inter-sectoral secondments, i.e., from academic to non-academic partners and vice versa). We have defined the following 11 cutting-edge research projects to be executed in three years: (1) explaining black-box models in terms of grey-box twin models; (2) from grey-box models to explainable models; (3) explaining Bayesian Networks; (4) explaining logical formulas; (5) multi-modal semantic grounding and model transparency; (6) explainable models for text production; (7) argumentation-based multi-agent recommender system; (8) customized interactive argumentation schemes for XAI; (9) personalized explanations by virtual characters; (10) interactions to mitigate human biases; and (11) explaining contracts and documentation of assets for companies.

Each single project is developed by an ESR with the guidance of his/her supervisors. Nevertheless, all ESRs contribute to create the framework depicted in Fig. 1, where ESRs are grouped in agreement with their research challenges. They all share a common open source software repository and collaborate in solving practical use cases posed by the companies involved in the project, regarding varied applications domains such as e-Health, education or telecommunications. Moreover, ESRs will do significant experimental work with human participants to inform and evaluate their algorithms. Accordingly, they will follow the Ethics guidelines for Trustworthy AI issued by the High-Level Expert Group on AI set up by the European Commission [8,9]. This will lead to generalizable methodologies and guidelines for generating and evaluating explanations.

The work developed by each ESR will end up with the publication of a PhD dissertation. Expected results are publications in top journals and conferences, but also additional resources such as open source software. Accordingly, each ESR has a personal career development plan with three main goals: (1) to develop the knowledge and skills required for performing high-quality doctoral research; (2) to develop transferable skills to enhance their personal effectiveness, leadership, management skills, research support, career and personal development, technology transfer and entrepreneurship, ethics and languages; and (3) to acquire ample insights in generating language and interactive solutions for XAI. To fulfil these training goals, the NL4XAI project includes a variety of network-wide training courses, motivating interactions at multiple levels (local,

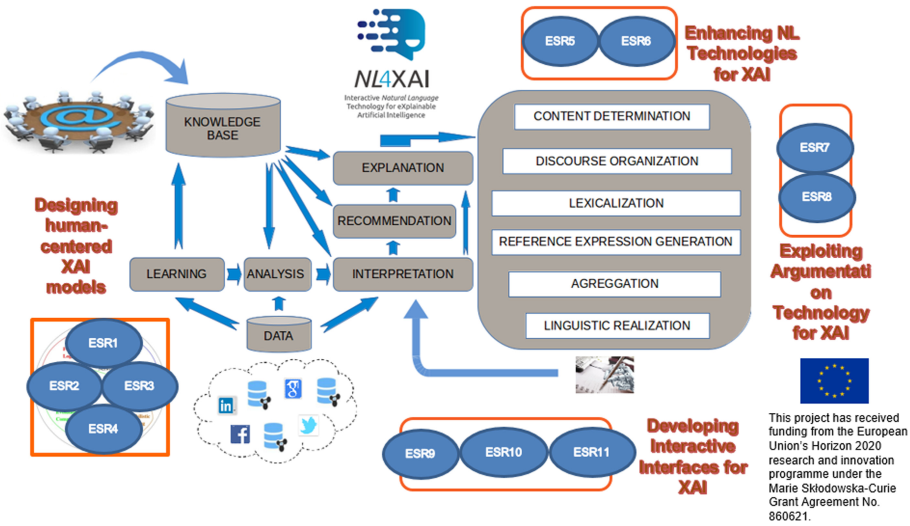

Fig. 1. NL4XAI framework for generating interactive explanations.

network-global) and at different depths (basics-general, specialisation-specific). All in all, we promote that ESRs collaborate actively during training activities and secondments (regarding both online and in person events).

4 Final Remarks

With the aim of generating narrative explanations in NL, we have identified the following major challenges [27] that will be jointly addressed by all ESRs:

- **Evaluation**: Develop "cheap but reliable" ways of estimating scrutability, trust, etc. Fair universal evaluation metrics and protocols supported by statistics for XAI are missing. Correlation among data-driven metrics and both intrinsic and extrinsic human evaluation needs to be studied.
- **Data Quality**: Develop techniques to let users know how results are influenced by data issues. Research on data bias is particularly encouraged.
- **Explanation effectiveness:** Develop models that make explicit how the effectiveness of an explanation depends on such factors as: the domain, the user, the context, the degree of interactivity, the level of precision and detail of the explanation, as well as on many concrete presentational choices involving the form and content of the explanation.

Finally, we note that ESRs will pay attention not only to technical but also to ethical and legal issues due to the worldwide social impact of XAI [8,9]. For example, Floridi et al. [14] defined an ethical framework for AI in which the concept of "explicability" captures the need for transparency and for accountability with reference to both "intelligibility" and "explainability". The interested reader is referred to the NL4XAI website for further insights on the development of this challenging project where the interplay between XAI and NL technologies leverages European innovation capacity.

Acknowledgment. The NL4XAI project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 860621.

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM, New York (2018). https://doi.org/10.1145/3173574.3174156
2. Alonso, J.M., Bugarín, A.: ExpliClas: automatic generation of explanations in natural language for Weka classifiers. In: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (2019). https://doi.org/10.1109/FUZZ-IEEE.2019.8859018
3. Alonso, J.M., Castiello, C., Magdalena, L., Mencar, C.: Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems. Studies in Computational Intelligence. Springer (2021). https://doi.org/10.1007/978-3-030-71098-9
4. Alonso, J.M., Ramos-Soto, A., Reiter, E., van Deemter, K.: An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems. In: Proceedings of the IEEE International Conference on Fuzzy Systems (2017). https://doi.org/10.1109/FUZZ-IEEE.2017.8015489
5. Alonso, J.M., Toja-Alamancos, J., Bugarín, A.: Experimental study on generating multi-modal explanations of black-box classifiers in terms of gray-box classifiers. In: Proceedings of the IEEE World Congress on Computational Intelligence (2020). https://doi.org/10.1109/FUZZ48607.2020.9177770
6. Budzynska, K., Villata, S.: Argument mining. IEEE Intell. Inform. Bull. **17**, 1–7 (2016)
7. Demollin, M., Shaheen, Q., Budzynska, K., Sierra, C.: Argumentation theoretical frameworks for explainable artificial intelligence. In: Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG). Dublin, Ireland (2020). https://www.aclweb.org/anthology/2020.nl4xai-1.10/
8. EU High Level Expert Group on AI: AI Ethics Guidelines for Trustworthy AI. Technical report, European Commission, Brussels, Belgium (2019). https://doi.org/10.2759/346720
9. EU High Level Expert Group on AI: The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. Technical report, European Commission, Brussels, Belgium (2019). https://doi.org/10.2759/002360
10. European Commission: Artificial Intelligence for Europe. Technical report, European Commission, Brussels, Belgium (2018). https://ec.europa.eu/digital-single-market/en/news/communicationartificial-intelligence-europe. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions (SWD(2018) 137 final)

11. Faille, J., Gatt, A., Gardent, C.: The natural language pipeline, neural text generation and explainability. In: Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG), Dublin, Ireland (2020). https://www.aclweb.org/anthology/2020.nl4xai-1.5/
12. Ferreira, T.C., van der Lee, C., van Miltenburg, E., Krahmer, E.: Neural data-to-text generation: a comparison between pipeline and end-to-end architectures. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, pp. 552–562. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1052
13. Fisher, W.R.: Human Communication as Narration: Toward a Philosophy of Reason, Value, and Action. University of South Carolina Press, Columbia (1989)
14. Floridi, L., et al.: AI4People - an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
15. Forrest, J., Sripada, S., Pang, W., Coghill, G.: Towards making NLG a voice for interpretable machine learning. In: Proceedings of the International Conference on Natural Language Generation (INLG), The Netherlands, pp. 177–182. Association for Computational Linguistics, Tilburg University (2018). https://doi.org/10.18653/v1/W18-6522
16. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: core tasks, applications and evaluation. J. Artif. Intell. Res. **61**, 65–170 (2018). https://doi.org/10.1613/jair.5477
17. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2018). https://doi.org/10.1145/3236009
18. Hennessy, C., Bugarin, A., Reiter, E.: Explaining Bayesian Networks in natural language: state of the art and challenges. In: Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG), Dublin, Ireland (2020). https://www.aclweb.org/anthology/2020.nl4xai-1.7/
19. Mariotti, E., Alonso, J.M., Gatt, A.: Towards harnessing natural language generation to explain black-box models. In: Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG), Dublin, Ireland (2020). https://www.aclweb.org/anthology/2020.nl4xai-1.6/
20. Mayn, A., van Deemter, K.: Towards generating effective explanations of logical formulas: challenges and strategies. In: Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG), Dublin, Ireland (2020). https://www.aclweb.org/anthology/2020.nl4xai-1.9/
21. Moryossef, A., Goldberg, Y., Dagan, I.: Improving quality and efficiency in plan-based neural data-to-text generation. In: Proceedings of the International Conference on Natural Language Generation (INLG), Tokyo, Japan, pp. 377–382. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/w19-8645
22. Narayan, S., Gardent, C.: Deep learning approaches to text production. In: Synthesis Lectures on Human Language Technologies, vol. 13, no. 1, pp. 1–199 (2020)
23. Parliament and Council of the European Union: General data protection regulation (GDPR) (2016). http://data.europa.eu/eli/reg/2016/679/oj

24. Pereira-Fariña, M., Bugarín, A.: Content determination for natural language descriptions of predictive Bayesian Networks. In: Proceedings of the Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), pp. 784–791. Atlantis Press (2019)
25. Polanyi, M.: The Tacit Dimension. Doubleday & Company Inc., New York (1966)
26. Rago, A., Cocarascu, O., Toni, F.: Argumentation-based recommendations: fantastic explanations and how to find them. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 1949–1955 (2018). https://doi.org/10.24963/ijcai.2018/269
27. Reiter, E.: Natural language generation challenges for explainable AI. In: Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI), pp. 3–7. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/W19-8402
28. Rieger, A., Theune, M., Tintarev, N.: Toward natural language mitigation strategies for cognitive biases in recommender systems. In: Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG), Dublin, Ireland (2020). https://www.aclweb.org/anthology/2020.nl4xai-1.11/
29. Sevilla, J.: Explaining data using causal Bayesian Networks. In: Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) at the International Conference on Natural Language Generation (INLG), Dublin, Ireland (2020). https://www.aclweb.org/anthology/2020.nl4xai-1.8/
30. Sierra, C., de Mántaras, R.L., Simoff, S.J.: The argumentative mediator. In: Proceedings of the European Conference on Multi-Agent Systems (EUMAS) and the International Conference on Agreement Technologies (AT), Valencia, Spain, pp. 439–454 (2016). https://doi.org/10.1007/978-3-319-59294-7_36
31. Stent, A., Bangalore, S.: Natural Language Generation in Interactive Systems. Cambridge University Press, Cambridge (2014)
32. Stepin, I., Alonso, J.M., Catala, A., Pereira, M.: Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In: Proceedings of the IEEE World Congress on Computational Intelligence (2020). https://doi.org/10.1109/FUZZ48607.2020.9177629
33. Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021). https://doi.org/10.1109/ACCESS.2021.3051315
34. Tintarev, N., Masthoff, J.: Explaining recommendations: design and evaluation. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 353–382. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_10
35. Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press, Cambridge (2008)
36. Williams, S., Reiter, E.: Generating basic skills reports for low-skilled readers. Nat. Lang. Eng. **14**, 495–535 (2008). https://doi.org/10.1017/S1351324908004725