Published in :
Calvaresi D., Najjar A., Winikoff M., Främling K. (eds) Explainable and Transparent AI and Multi-Agent Systems. EXTRAAMAS 2021. Lecture Notes in Computer Science, vol 12688. Springer, Cham. https://doi.org/10.1007/978-3-030-82017-6_20
which should be cited to refer to this work.

EXPECTATION: Personalized Explainable Artificial Intelligence for Decentralized Agents with Heterogeneous Knowledge

Davide Calvaresi[✉4][0000−0001−9816−7439], Giovanni Ciatto[1][0000−0002−1841−8996], Amro Najjar[2][0000−0001−7784−6176], Reyhan Aydoğan[3][0000−0002−5260−9999], Leon Van der Torre[2][0000−0003−4330−3717], Andrea Omicini[1][0000−0002−6655−3869], and Michael Schumacher[4][0000−0002−5123−5075]

[1] ALMA MATER STUDIORUM – Università di Bologna, Cesena, Italy
{giovanni.ciatto, andrea.omicini}@unibo.it
[2] University of Luxembourg, Luxembourg
{amro.najjar, leon.vandertorre}@uni.lu
[3] Özyeğin University, Istanbul, Turkey
reyhan.aydogan@ozyegin.edu.tr
[4] University of Applied Sciences and Arts Western Switzerland HES-SO, Switzerland
{davide.calvaresi, michael.schumacher}@hevs.ch

Abstract. Explainable AI (XAI) has emerged in recent years as a set of techniques and methodologies to interpret and explain machine learning (ML) predictors. To date, many initiatives have been proposed. Nevertheless, current research efforts mainly focus on methods tailored to specific ML tasks and algorithms, such as image classification and sentiment analysis. However, explanation techniques are still embryotic, and they mainly target ML experts rather than heterogeneous end-users. Furthermore, existing solutions assume data to be centralised, homogeneous, and fully/continuously accessible—circumstances seldom found altogether in practice. Arguably, a system-wide perspective is currently missing.

The project named "Personalized Explainable Artificial Intelligence for Decentralized Agents with Heterogeneous Knowledge" (EXPECTATION) aims at overcoming such limitations. This manuscript presents the overall objectives and approach of the EXPECTATION project, focusing on the theoretical and practical advance of the state of the art of XAI towards the construction of *personalised* explanations in spite of *decentralisation* and *heterogeneity* of knowledge, agents, and explainees (both humans or virtual).

To tackle the challenges posed by personalisation, decentralisation, and heterogeneity, the project fruitfully combines abstractions, methods, and approaches from the multi-agent systems, knowledge extraction / injection, negotiation, argumentation, and symbolic reasoning communities.

Keywords: Multi-agent systems · eXplanable AI · CHIST-ERA IV · Personalisation · Decentralisation · EXPECTATION

1 Background and Motivations

In recent decades, data-driven decision-making processes have increasingly influenced strategic choices. This applies to both virtual and humans' decisional needs. The application domains of Machine learning (ML) algorithms are broadening [1,2]. Ranging from finance to healthcare, ML supports humans in making informed decisions based on the information buried within enormous amounts of data. However, most effective ML methods are inherently *opaque*, meaning that it is hard for humans (if possible at all) to grasp the reasoning *hidden* in their predictions (so-called black boxes). To mitigate the issues arising from such opaqueness, several techniques and methodologies aiming at inspecting ML models and predictors have been proposed under the eXplainable Artificial Intelligence (XAI) umbrella [3,4] (e.g., feature importance estimators, rule lists, and surrogate trees [5]). Such tools enable humans to understand, inspect, analyse – and therefore trust – the operation and outcomes of AI systems effectively.

The many XAI-related initiatives proposed so far constitute the building blocks for making tomorrow's intelligent systems explainable and trustable. However, to date, the ultimate goal of letting intelligent systems provide not only valuable recommendations but also *motivations* and *explanations* for their suggestions – possibly, interactively – is still unachieved. Indeed, current research efforts focus on specific methods and algorithms, often tailored to single ML tasks—e.g. classification and, in particular, image classification. For instance, virtually all approaches proposed so far target supervised learning, and in particular, classification tasks [6,3,4]—and many of them are tailored on neural networks [7]. In other words, there is still a long way to *generality* [8].

Moreover, while existing XAI solutions do an excellent job on inspecting ML algorithms, current interpretation/explanations provide valuable insights solely profitable by human *experts*, entirely neglecting the need for producing more broadly accessible or personalised explanations that everybody could understand. Recalling their social nature, explanations should rather be *interactive* and tailored on the explainee's cognitive capabilities and background knowledge to be effective [9,10].

To complicate this matter, existing XAI solutions assume data to be centralised, homogeneous, and fully/continuously available for operation [8]. Such circumstances rarely occur in real-world scenarios. For example, data is often scattered through many administrative domains. Thus, even when carrying similar information, datasets are commonly structured according to different schemas—when not lacking structure at all. Privacy and legal constraints complete the picture by making it unlikely for data to be fully available at any given moment. In other words, the availability of data is more frequently *partial* rather than total. Therefore, explainable intelligent systems should be able to deal with scattering, decentralisation, heterogeneity, and unavailability of data, rather than requiring data to be centralised and standardised before even starting to process it—which would impose heavy technical, administrative, and legal constraints on the production of both recommendations and explanations.

Summarising, further research is needed to push XAI towards the construction of *personalised* explanations, which can be built in spite of *decentralisation* and *heterogeneity* of information—possibly, out of the interaction among intelligent software systems and human or virtual explainees.

Clearly, tackling personalisation, decentralisation, and heterogeneity entails challenges from several perspectives. On the one hand, personalisation of explanations must cope with the need for providing human-intelligible (i.e., *symbolic*) explanations of incremental complexity, possibly *iteratively* adapting to the cognitive capabilities, and background knowledge of the users who are receiving the explanation. In turn, it requires enabling an *interactive* explanation process both within the intelligent systems themselves (i.e., agent to agent) and with the end-users. On the other hand, decentralisation of data opens to questioning how explanations can be produced or aggregated without letting data cross administrative borders. Therefore, the need for *collaboration* among multiple cross-domain software entities is imperative. Finally, the challenge of heterogeneity, of both data and ML techniques used to mine information out of it, dictates the detection of some *lingua franca* to present recommendations and explanations to the users in intelligible forms.

To address these challenges, the EXPECTATION project has been recently recommended for funding – along with other 11 projects – as part of the CHIST-ERA 2019 call[5] concerning "Explainable Machine Learning-based Artificial Intelligence". The project has started on April 1, 2021 and it will last up to the of March 2024. In the remainder of this paper, we discuss how the project plans to tackle the challenges posed by personalisation, decentralisation, and heterogeneity, by fruitfully combining abstractions, methods, and approaches from the multi-agent systems, knowledge extraction/injection, negotiation, argumentation, and symbolic reasoning research areas.

2 State of the Art

The generation of personalised explanation for decentralised and heterogeneous intelligent agents roots in several disciplines, including XAI, agreement technologies, personalisation, and AI ethics.

2.1 Explainable Agency

Neuro-symbolic integration [11,12] aims at bridging the gap between symbolic and sub-symbolic AI, reconciling the two key branches of AI (connectionist AI – relying on connectionist networks inspired from human neurons, and symbolic AI – relying on logic, symbols, and reasoning) [13]. Sub-symbolic techniques (e.g., pattern recognition and classification) can offer excellent performance. However, their outcomes can be biased and difficult to understand (if possible at all). Seeking trust, transparency, and the possibility to debug sub-symbolic predictors (so-called black boxes), the XAI community relies on reverse engineering

[5] `https://www.chistera.eu/projects-call-2019`

models trained on unknown datasets generating plausible explanations fitting the outcome produced by the black box [14]. A typical practice is to train an interpretable machine learning model (e.g., decision trees, linear model, or rules) with the outcome of a black box [3,15,16].

Explainable agents go beyond the mere application of sub-symbolic ML mechanisms. Agents can leverage symbolic AI techniques (e.g., logic and planning languages), which are easier to trace, reason about, understand, debug, and explain [17]. However, they can still partially rely on ML predictors, thus deeming necessary to be explaining their overall behavior (relying on neuro-symbolic integration). Endowing virtual agents with explanatory abilities raises trust, acceptability, and reduces possible failures due to misunderstandings [14,18]. Yet, it necessary to consider user characterisation (e.g., age, background, and expertise), the context (e.g., why do the user need the explanation), and the agents' limits [14].

Built-in explainability is still rare in literature. Most of the works utterly provide indicators which "should serve" as an explanation for the human user [3]. To date, such approaches have been unable to produce satisfying human-understandable explanations. Nevertheless, more recent contributions employ neuro-symbolic integration to identifying factors influencing the human comprehension of representation formats and reasoning approaches [19].

2.2 Agreement Technologies

Understanding other parties' interests and preferences is crucial in human social interaction. It enables the proposal of reasonable bids to resolve conflicts effectively [20,21]. *Agreement technologies* (AT) [22] literature counts several techniques to automatically learn, reproduce, and possibly predict an opponent's preferences and bidding strategies in conflict resolution scenarios [23].

AT are mostly based on heuristics [24,25] and traditional ML methods (e.g., decision trees [26,27], Bayesian learning [28,29,30], and concept-based learning [31,32]) and rely on possibly numerous bid exchanges regulated by negotiation protocols [33]. By exploiting such techniques, machines can negotiate with humans seamlessly, resolving conflicts with a high degree of mutual understanding [34]. Nevertheless, in human-agent negotiation, the complexity skyrockets. Humans leverage on semantic and reasoning (e.g., employing similarities/differences) while learning about the competitors' preferences and generating well-targeted offers. Conversely to agent-agent, the number of exchanged bids between parties is limited due to the nature of human interactions, and may employ unstructured data. Therefore, classical opponent modeling techniques used in automated negotiation in which thousands of bids are exchanged may not be suitable, and additional reasoning to understand humans' intentions, interests, arguments, and explanations supporting their proposals is required [35,36]. To the best of our knowledge, there is no study incorporating exchanged arguments or *explanations* into opponent modeling in agent-based negotiation literature.

Without explanations, human users may attribute a wrong *state of mind* to agents/robots [18]. Thus, the creation of an effective agent-based explainable

AT for human-agent interactions and the realisation of a common understanding would require the integration of *(i)* ontology reasoning, *(ii)* understanding humans' preferences/interests by reasoning on any type of information provided during the negotiation, and *(iii)* generating well-targeted offers with their supportive explanations or motivations (i.e., why the offer can be acceptable for their human counterpart). To the best of our knowledge, the state of the art still needs concrete contributions concerning the three directions mentioned above. Moreover, albeit the need for personalised motivations and arguments (e.g., considering user expertise, personal attributes, and goals) is well known in literature [14], most of the existing works are rather conceptual and do not consider the overall big picture [37]. Furthermore, no work addresses explanation personalisation in the context of heterogeneous systems combining sub-symbolic (e.g., neural network) and symbolic (agents/robots) AI mechanisms.

2.3 AI Ethics

Due to the growing adoption of intelligent systems, machine ethics and AI ethics have received a deserved increasing attention from scientists working in various domains [38]. The growing safety, ethical, societal, and legal impacts of AI decisions are the main reason behind this surge of interest [39]. In literature, AI ethics includes implicitly- and explicitly-moral agents. In both cases, intelligent systems depend on human intervention to distinguish moral from immoral behaviour. However, on the one hand, implicitly-moral agents are ethically constrained from having immoral behaviour via rules set by the human designer [38]. On the other hand, explicitly-ethical agents (or agents with functional morality) presume to be able to morally judge themselves (having guidelines or examples of what is good and bad [38]).

Summarising, AI systems can have implicit and explicit ethical notions. The main advantage of implicit AI ethics is that they are simple to develop and control, being incapable of unethical behaviour. Nevertheless, this simplicity implies mirroring the ethic standing point and perception of the designer. Explicit-ethics systems affirm to autonomously evaluate the normative status of actions and reason independently about what they consider unethical, thus being able to solve normative conflicts. Furthermore, they could bend/violate some rules, resulting in better fulfilment of overarching ethical objectives. However, the main shortcoming of these systems is their complexity and possible unexpected behaviour.

3 The EXPECTATION Approach

This section elaborates on the limitations elicited from the state of art, the related challenges, and formalises the needed interventions. The six major limitations identified are:

(L1) Opaqueness of sub-symbolic predictors. Most ML algorithms leverage a sub-symbolic representation of knowledge that is hard to debug for experts

and hard to interpret for common people. Thus, the compliance of internal mechanisms and results with ethical principles and regulations cannot be verified/ensured.

(L2) **Heterogeneity of rule extraction techniques**. Extracting general-purpose symbolic rules from any sort of sub-symbolic predictor can be a difficult task (if possible, at all). Indeed, the nature of the data and the particular predictor at hand significantly impact the quality (i.e., the intelligibility) of the extracted rules. Furthermore, existing techniques to extract rules to produce explanations mostly leverage structured, low-dimensional data, given the scarcity of methods supporting more complex data (i.e., images, videos, or audios). In particular, most of the existing works interpreting sub-symbolic mechanisms place interpretable mechanisms (i.e., decision-tree) on top of the predictors, thereby interpreting (e.g., reconstructing) from outside their outcomes without really mirroring their internal mechanisms.

(L3) **Manual amending and integration of heterogeneous predictors**. The update and integration of already pre-trained predictors are usually handcrafted and poorly automatable. Moreover, it heavily relies on datasets that might be available only for a limited period. Therefore, a sustainable, automatable, and seamless sharing/reusing/integrating of knowledge from diverse predictors is still unsatisfactory.

(L4) **Lack of personalisation**. Current XAI approaches are mostly one-way processes (e.g., interactive interactions are rarely involved) and do not consider the explainee's context and background. Thus, the customisation and personalisation of the explanations are still open challenges.

(L5) **Tendency of centralisation in data-driven AI**. The development of sub-symbolic predictors usually involves the centralisation of training data in a single point, which raises privacy concerns. Thus, letting a system composed of several distributed intelligent components learning without centralising data is still an open challenge.

(L6) **Lack of explanation integration in Agreement Technologies**. Current negotiation and argumentation frameworks mostly leverage well-structured interactions and clearly defined objectives, resources, and goals. Current AT are not suitable for providing interactive explanations nor for reconciling fragmented knowledge. Moreover, although a few works explored more sophisticated mechanisms (e.g., adopting semantic similarities via subsumption to relate alternative values within a single bid), the need for ontological reasoning to infer the relationship between several issues – possibly pivotal in negotiation and argumentation of explanations – is still unmet.

To overcome the limitation mention, EXPECTATION formalises the following objectives:

(O1) To define an agent-based model embedding ML predictors relying on heterogeneous (though potentially similar/complementary) knowledge, as in training datasets, contextual assumptions & ontologies.

(O2) To design and implement a decentralised agent architecture capable of integrating symbolic knowledge and explanations produced by individual agents.

(O3) To define and implement agent strategies for cooperation, negotiation, and trust establishment for providing personalised explanations according to the user context.

(O4) To investigate, implement, and evaluate multi-modal explanation communication mechanisms (visual, auditory, cues, etc.), the role of the type of agent providing these explanations (e.g., robot, virtual agents), and their role in explanation personalisation.

(O5) To validate and evaluate the personalised explainability results, as well as the agent-based XAI approach for heterogeneous knowledge, within the context of a prototype, focused on food and nutrition recommendations.

(O6) To investigate the specific ethical challenges that XAI is able to meet and when and to what extent explicability is legally required in European regulations, considering the AI guidelines and evaluation protocols published by the national and European institutions (e.g., the Data Protection Impact Analysis thanks to the open-source software PIA, CNIL guidelines), as well as recent research on the ethics of recommender systems w.r.t. values such as transparency and fairness.

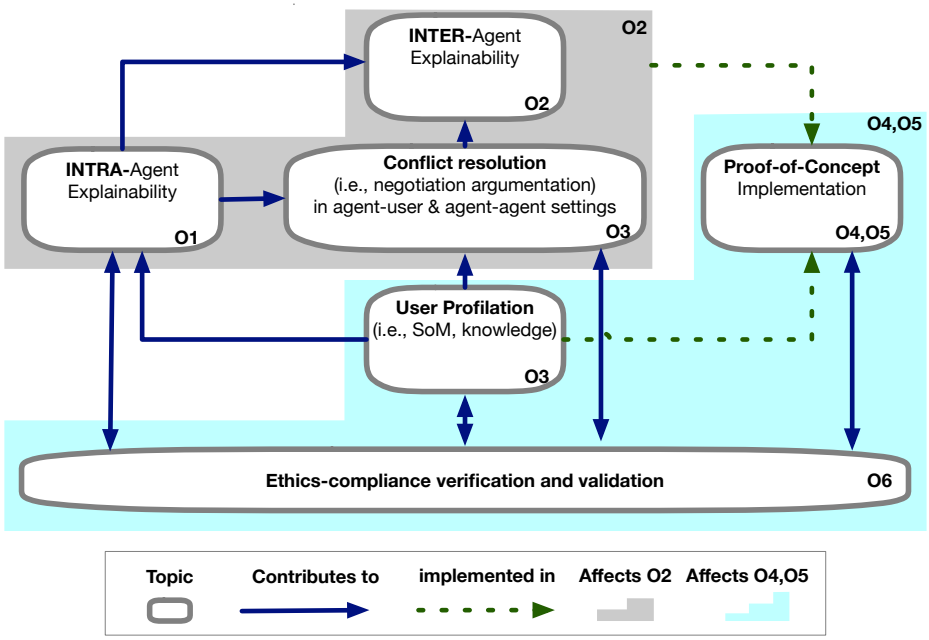

Fig. 1. EXPECTATION's objectives, topics, and respective interconnections.

The aforementioned objectives are clearly interdependent. In particular, Figure 1 groups and organises the objectives per contribution, effect, and implementation among each other.

3.1 Research Method

Despite being still in its early stage, the project's roadmap has already been established. EXPECTATION's research and development activities will be carried out along two orthogonal dimensions – namely *intra*- and *inter*-agent ones –, as depicted in Figure 2.

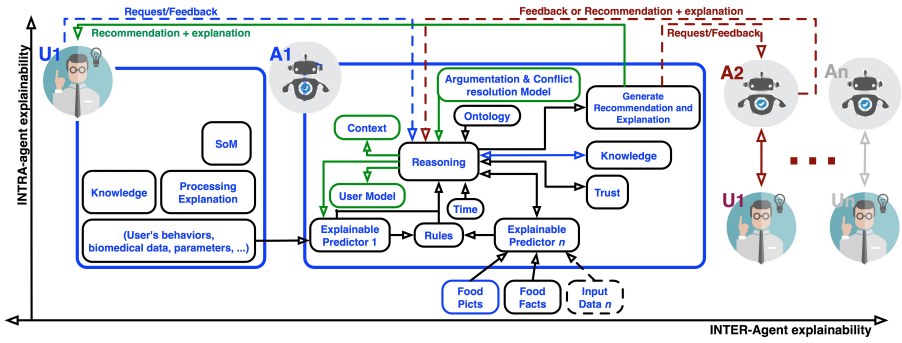

Fig. 2. Main components and interactions of the proposed architecture.

The envisioned scenario for this project assumes a 1-to-1 mapping between end-users and software agents (cf. Figure 2, rightmost part). Therefore, each software agent interacts with a single user in order to *(i)* acquire their contextual data (cf. blue dashed line in Figure 2), and *(ii)* provide them with personalised explanations taking that contextual information into account (cf. green solid line in Figure 2). This is the purpose of what we call *intra*-agent explainability.

However, the idea of building agents that provide precise recommendations by solely leveraging on the data acquired from a single user is unrealistic. Accordingly, we envision agents to autonomously debate and negotiate with each other to mutually complement and globally improve their knowledge, thus generating personalised and accurate recommendations. Addressing this challenge is the purpose of what we call *inter*-agent explainability.

On the one hand, intra-agent explainability focuses on deriving explainable information at the local level – where contextual information about the user is most likely available – and on presenting it to the user in a personalised way. To do so, symbolic knowledge extraction and injection play a crucial role. The former lets agents fully exploit the predictive performance of conventional ML-based black-box algorithms while still enabling the production of intelligible information to be used for building personalised explanations. Conversely, by injecting symbolic knowledge in ML-based systems, agents will be able to update, revise, and correct the functioning of ML-based predictors by taking into account users' contextual information and feedback.

On the other hand, inter-agent explainability focuses on enabling the agents to exploit negotiation and argumentations to mutually improve their predictive

capabilities by exchanging the symbolic knowledge they have extracted from given black boxes. Even in this context, the role of symbolic knowledge extraction is of paramount importance as it enables exchanges of aggregated knowledge coming from different ML-predictors—which possibly offer different perspectives on the problem at hand. To this end, inter-agent explainability requires formalising interaction protocols specifying what actions are possible and how to represent this information so that both parties can understand and interpret it seamlessly. Moreover, inter-agent interactions will require reasoning mechanisms handling heterogeneous data received from other agents, including techniques to detect conflicts and adopt resolution or mitigation policies accordingly.

By combining intra- and inter-agent explainability, EXPECTATION will be able to tackle decentralisation (of both data and agents), heterogeneity (of both data and analysis techniques), and users' privacy simultaneously. Indeed, the proposed approach does not require data to be centralised to allow training and knowledge extraction. Therefore, each agent can autonomously take care of the local data it has access to by exploiting the ML-based analysis technique it prefers, while joint learning is delegated to decentralised negotiation protocols which only exchange aggregated knowledge. Users' personal data is expected to remain close to the user, while agents are in charge of blending the extracted symbolic knowledge with the general-purpose background knowledge jointly attained by the multi-agent systems via negotiation and argumentation. Heterogeneity is addressed indirectly via knowledge extraction, which provides a *lingua franca* for knowledge sharing in the form of logic facts and rules.

Notably, knowledge extraction is what enables bridging *intra-* and *inter*-agent explainability too, as it enables the exchange of the extracted knowledge via negotiation and argumentation protocols—which already rely on the exchange of symbolic information.

Knowledge injection closes the loop by letting the knowledge acquired via interaction to be used to improve the local data and analytic capabilities of each individual agent. Finally, the purposes of preserving privacy and complying with ethical implications are addressed by only allowing agents to share aggregated symbolic knowledge. Moreover, we envision to equip the agents with ethics reasoning engines combining techniques from both implicit and explicit ethics.

4 Discussion

To test the advancement produced by EXPECTATION, we envision combining the techniques mentioned above in a proof of concept cantered on a topic which nowadays is delicate more than ever: a nutrition recommender system, fostering a responsible and correct alimentation. Such a prototype will be tested and evaluated according to the user-subjective such as understandability, trust, acceptability, soundness, personalisation, perceived system autonomy, perceived user autonomy, and fairness. The envisioned agent-based recommender system is intended to operate as a virtual assistant equipped with personalised explanatory capabilities. This would make it possible to tackle two dimensions of the quest

for a correct regime *(i)* trust and acceptance, and *(ii)* autonomous personalisation, education, and explicability. In particular, the user will be provided with transparent explanations about the recommendation received. The purpose of the explanations is multi-faceted: *(i)* educative (i.e., improve the user knowledge and raising his/her awareness about a given topic/suggestion), *(ii)* informative (i.e., indicate the user on how the system works), and *(iii)* motivational (i.e., it helps the user understanding how personal characteristics and decisions lead to favorable/adverse outcomes).

Overall, EXPECTATION is expected to impact beyond its lifespan. Such an impact encompasses several aspects and is four-folded.

Impact of theoretical outcomes. Production of mechanisms to extract, combine, explain, negotiate heterogeneous symbolic knowledge as well as cooperation and negotiation strategies.

Impact of technological outcomes. Fostering the adoption of intelligent systems in health and safety-critical domains and inspiring new technology leveraging innovative multi-modal explanation communication mechanisms.

Impact in application domains. We expect uptake of the project results in sectors (commercial/academic) such as eHealth, prevention, wellbeing applications, and distribution and restoration.

Impact of ethical aspects. Given the sensitive nature of personal data in the context of the project, the proposed XAI prototype will develop generalisable mechanisms to ensure compliance, fairness, transparency, and trust.

Acknowledgments

This work has been partially supported by the CHIST-ERA grant CHIST-ERA-19-XAI-005, and by *(i)* the Swiss National Science Foundation (G.A. 20CH21_195530), *(ii)* the Italian Ministry for Universities and Research, *(iii)* the Luxembourg National Research Fund (G.A. INTER/CHIST/19/14589586), *(iv)* the Scientific and Research Council of Turkey (TÜBİTAK, G.A. 120N680).

References

1. Zubair Md Fadlullah, Fengxiao Tang, Bomin Mao, Nei Kato, Osamu Akashi, Takeru Inoue, and Kimihiro Mizutani. State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Communications Surveys Tutorials*, 19(4):2432–2455, 2017.
2. Dirk Helbing. Societal, economic, ethical and legal challenges of the digital revolution: From big data to deep learning, artificial intelligence, and manipulative technologies. In *Towards Digital Enlightenment. Essays on the Dark and Light Sides of the Digital Revolution*, pages 47–72. Springer, 2019.
3. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42, 2019.

4. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58(December 2019):82–115, 2020.
5. Roberta Calegari, Giovanni Ciatto, and Andrea Omicini. On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale*, 14(1):7–32, 2020.
6. Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. In Karolj Skala, Marko Koricic, Tihana Galinac Grbac, Marina Cicin-Sain, Vlado Sruk, Slobodan Ribaric, Stjepan Gros, Boris Vrdoljak, Mladen Mauher, Edvard Tijan, Predrag Pale, and Matej Janjic, editors, *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018)*, pages 210–215, Opatija, Croatia, 21–25 May 2018. IEEE.
7. Evren Dağlarli. Explainable artificial intelligence (xAI) approaches and deep meta-learning models. In Marco Antonio Aceves-Fernandez, editor, *Advances and Applications in Deep Learning*, chapter 5. IntechOpen, London, UK, 2020.
8. Giovanni Ciatto, Roberta Calegari, Andrea Omicini, and Davide Calvaresi. Towards XMAS: eXplainability through Multi-Agent Systems. In Claudio Savaglio, Giancarlo Fortino, Giovanni Ciatto, and Andrea Omicini, editors, *AI&IoT 2019 – Artificial Intelligence and Internet of Things 2019*, volume 2502 of *CEUR Workshop Proceedings*, pages 40–53. Sun SITE Central Europe, RWTH Aachen University, November 2019.
9. Giovanni Ciatto, Michael I. Schumacher, Andrea Omicini, and Davide Calvaresi. Agent-based explanations in AI: Towards an abstract framework. In Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling, editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, volume 12175 of *Lecture Notes in Computer Science*, pages 3–20. Springer, Cham, 2020. 2$^{\text{nd}}$ International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers.
10. Giovanni Ciatto, Davide Calvaresi, Michael I. Schumacher, and Andrea Omicini. An abstract framework for agent-based explanations in AI. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1816–1818, Auckland, New Zeland, May 2020. International Foundation for Autonomous Agents and Multiagent Systems. Extended Abstract.
11. Giuseppe Pisano, Giovanni Ciatto, Roberta Calegari, and Andrea Omicini. Neuro-symbolic computation for XAI: Towards a unified model. In Roberta Calegari, Giovanni Ciatto, Enrico Denti, Andrea Omicini, and Giovanni Sartor, editors, *WOA 2020 – 21th Workshop "From Objects to Agents"*, volume 2706 of *CEUR Workshop Proceedings*, pages 101–117, Aachen, Germany, October 2020. Sun SITE Central Europe, RWTH Aachen University. Bologna, Italy, 14–16 September 2020.
12. Benedikt Wagner and Artur d'Avila Garcez. Neural-symbolic integration for fairness in AI. In Andreas Martin, Knut Hinkelmann, Hans-Georg Fill, Aurona Gerber, Doug Lenat, Reinhard Stolle, and Frank van Harmelen, editors, *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021)*, volume 2846 of *CEUR Workshop Proceedings*, Stanford University, Palo Alto, CA, USA, 22–24 March 2021. CEUR-WS.org.

13. Paul Smolensky. Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, 1(2):95–109, 1987.
14. Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *18^{th} International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'19)*, pages 1078–1088, Montreal, QC, Canada, 13–17 May 2019. International Foundation for Autonomous Agents and Multiagent Systems.
15. Roberta Calegari, Giovanni Ciatto, Jason Dellaluce, and Andrea Omicini. Interpretable narrative explanation for ML predictors with LP: A case study for XAI. In Federico Bergenti and Stefania Monica, editors, *WOA 2019 – 20^{th} Workshop "From Objects to Agents"*, volume 2404 of *CEUR Workshop Proceedings*, pages 105–112. Sun SITE Central Europe, RWTH Aachen University, Parma, Italy, 26–28 June 2019.
16. Robert Andrews, Joachim Diederich, and Alan B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6):373–389, December 1995.
17. Roberta Calegari, Giovanni Ciatto, Viviana Mascardi, and Andrea Omicini. Logic-based technologies for multi-agent systems: A systematic literature review. *Autonomous Agents and Multi-Agent Systems*, 35(1):1:1–1:67, 2021. Collection "Current Trends in Research on Software Agents and Agent-Based Software Development".
18. Thomas Hellström and Suna Bensch. Understandable robots - what, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1):110–123, 2018.
19. Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.
20. Tim Baarslag, Michael Kaisers, Enrico Gerding, Catholijn Jonker, and Jonathan Gratch. Computers that negotiate on our behalf: Major challenges for self-sufficient, self-directed, and interdependent negotiating agents. In *Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, Visionary Papers*, pages 143–163, 2017.
21. Catholijn M. Jonker, Reyhan Aydoğan, Tim Baarslag, Joost Broekens, Christian A. Detweiler, Koen V. Hindriks, Alina Huldtgren, and Wouter Pasman. An Introduction to the Pocket Negotiator: A General Purpose Negotiation Support System. In Natalia Criado Pacheco, Carlos Carrascosa, Nardine Osman, and Vicente Julián Inglada, editors, *Multi-Agent Systems and Agreement Technologies*, pages 13–27, Cham, 2017. Springer International Publishing.
22. Sascha Ossowski, editor. *Agreement Technologies*, volume 8 of *Law, Governance and Technology Series*. Springer Netherlands, 2012.
23. T. Baarslag, Mark J. C. Hendrikx, K. Hindriks, and C. Jonker. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30:849–898, 2015.
24. Reyhan Aydoğan, T. Baarslag, K. Hindriks, C. Jonker, and P. Yolum. Heuristics for using CP-nets in utility-based negotiation without knowing utilities. *Knowledge and Information Systems*, 45:357–388, 2014.
25. N. Jennings, Peyman Faratin, Alessio Lomuscio, Simon Parsons, Michael Wooldridge, and Carles Sierra. Automated negotiation: Prospects, methods and challenges. *Group Decision and Negotiation*, 10:199–215, March 2001.

26. Reyhan Aydoğan, Ivan Marsá-Maestre, Mark Klein, and Catholijn Jonker. A Machine Learning Approach for Mechanism Selection in Complex Negotiations. *Journal of Systems Science and Systems Engineering*, 27, 2018.
27. Litan Ilany and Ya'akov Gal. Algorithm selection in bilateral negotiation. *Autonomous Agents and Multi-Agent Systems*, 30(4):697–723, July 2016.
28. Koen V. Hindriks and Dmytro Tykhonov. Opponent modelling in automated multi-issue negotiation using bayesian learning. In Lin Padgham, David C. Parkes, Jörg P. Müller, and Simon Parsons, editors, *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, volume 1, pages 331–338, Estoril, Portugal, 12–16 May 2008. IFAAMAS.
29. Chao Yu, Fenghui Ren, and Minjie Zhang. *An Adaptive Bilateral Negotiation Model Based on Bayesian Learning*, volume 435, pages 75–93. Springer, January 2013.
30. Dajun Zeng and Katia Sycara. Bayesian learning in negotiation. *International Journal of Human-Computer Studies*, 48(1):125–141, 1998.
31. Reyhan Aydogan and Pinar Yolum. Ontology-based learning for negotiation. In *2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2009)*, volume 2, pages 177–184, January 2009.
32. Boris A. Galitsky, Sergei O. Kuznetsov, and Mikhail V. Samokhin. Analyzing conflicts with concept-based learning. In Frithjof Dau, Marie-Laure Mugnier, and Gerd Stumme, editors, *Conceptual Structures: Common Semantics for Sharing Knowledge*, pages 307–322. Springer Berlin Heidelberg, 2005.
33. Ivan Marsa-Maestre, Mark Klein, Catholijn M. Jonker, and Reyhan Aydoğan. From problems to protocols: Towards a negotiation handbook. *Decision Support Systems*, 60:39–54, 2014.
34. Yinon Oshrat, Raz Lin, and Sarit Kraus. Facing the challenge of human-agent negotiations via effective general opponent modeling. In *8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09)*, volume 1, pages 377–384. IFAAMAS, 2009.
35. Onat Güngör, Umut Çakan, Reyhan Aydoğan, and Pinar Özturk. Effect of awareness of other side's gain on negotiation outcome, emotion, argument, and bidding behavior. In Reyhan Aydoğan, Takayuki Ito, Ahmed Moustafa, Takanobu Otsuka, and Minjie Zhang, editors, *Recent Advances in Agent-based Negotiation*, pages 3–20, Singapore, 2021. Springer Singapore.
36. Philippe Pasquier, Ramon Hollands, Frank Dignum, Iyad Rahwan, and Liz Sonenberg. An empirical study of interest-based negotiation. *Autonomous Agents and Multi-Agent Systems*, 22:249–288, 2011.
37. Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 676–682, 2017.
38. James H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.
39. Davide Calvaresi, Michael Schumacher, and Jean-Paul Calbimonte. Personal data privacy semantics in multi-agent systems interactions. In Yves Demazeau, Tom Holvoet, Juan M. Corchado, and Stefania Costantini, editors, *Advances in Practical Applications of Agents, Multi-Agent Systems, and Trustworthiness. The PAAMS Collection - 18th International Conference (PAAMS 2020)*, volume 12092 of *Lecture Notes in Computer Science*, pages 55–67, L'Aquila, Italy, 7–9 October 2020. Springer.