The FATE System: FAir, Transparent and Explainable Decision Making

Joachim de Greeff [a], Maaike H.T. de Boer [a], Fieke H. J. Hillerström [a], Freek Bomhof [a], Wiard Jorritsma [a] and Mark Neerincx [a]

[a] *TNO, Anna van Buerenplein 1, 2595 DA, The Hague, The Netherlands*

Abstract

AI tools are becoming more commonly used in a variety of application domains. In this paper, we describe a system named FATE that combines state of the art AI tools. The goal of the FATE system is decision support with use of ongoing human-AI co-learning, explainable AI and fair, bias-free and secure usage of data. These topics are societally very relevant for the update of AI-based support systems, but the manner in which to bring these together into a working system is far from trivial. We describe the various functional components that comprise the system, share our experience with the set-up of such a system, explain how it can be used in a variety of use cases, taking into account multiple user roles. Finally, we reflect on this and provide an outlook for the continuation of this development.

Keywords

FAIR AI, Hybrid AI, Explainable AI, Bias, Secure Learning, Knowledge Engineering, Co-Learning

1. Introduction

AI tools are becoming more advanced and ubiquitous in a wide variety of application domains. Yet, as of now they tend to be fairly specialized and not necessarily tailored towards a large variety of (naïve) users. Moreover, the application of AI tools in a fair, bias-free and explainable manner remains a big challenge. The FATE (FAir, Transparent, Explainable) system aims to unify a set of AI tools addressing these issues into a ready-to-use decision support system that can be utilized by end users in various roles and domains.

FATE combines machine learning with knowledge engineering methods to establish human-AI co-learning for ongoing human-AI performance improvements. We develop AI capabilities to provide classifications, predictions, advice and decisions in a *fair, understandable-trustworthy*, *controllable* and *secure* manner [1]. Fairness is realized through *bias management* (i.e., prevention, identification and mitigation) [2]. Adequate human understanding and trust is realized by the capability to provide *personalized explanation*s [3] alongside the AI-generated output. Meaningful human control is supported by the capability to integrate (explicit) human knowledge into the ongoing learning process, i.e., *knowledge engineering*, for instance via the provision of new concepts in the input of the machine learning and by adding causal relations for explanations [4]. And finally, secure learning is realized through *federated learning* [20], which will dictate additional constraints on the other capabilities (i.e., bias management, personalized explanations, knowledge engineering).

The FATE system is conceived in a generic manner, meaning that the system is domain-agnostic. By applying it to a number of use cases in different domains, the system becomes populated with data, knowledge and user preferences. The currently targeted use cases are diabetes type 2 risk prediction and management (healthcare), industrial predictive maintenance (industry) and risk assessment of convicts' fleeing chance (justice and security).

The creation of the FATE system was a challenge, because combining the different AI research areas as well as maintaining a generic system has not yet been done before. The method of socio-cognitive engineering [5] is used to perform research and development of the FATE system in an incremental manner. In this paper we share our experience with creating such a system.

In the next section we describe the different components of the FATE system. Following that, we reflect on the process and experiences of the first year of working on the system and we conclude with an outlook towards future work.

2. Components of the FATE system

The FATE system is conceived as a generic decision support system, that can be applied within a variety of different domains and use cases. Through application to a specific use case, the FATE system becomes instantiated for this particular context. A working use case is for example decision support for diabetes type 2, in which the system can give health-related advice to both healthcare professionals and patients, tailored towards their respective contexts. Thus, the use case specifies the user roles, domain knowledge, data sources and advice type.

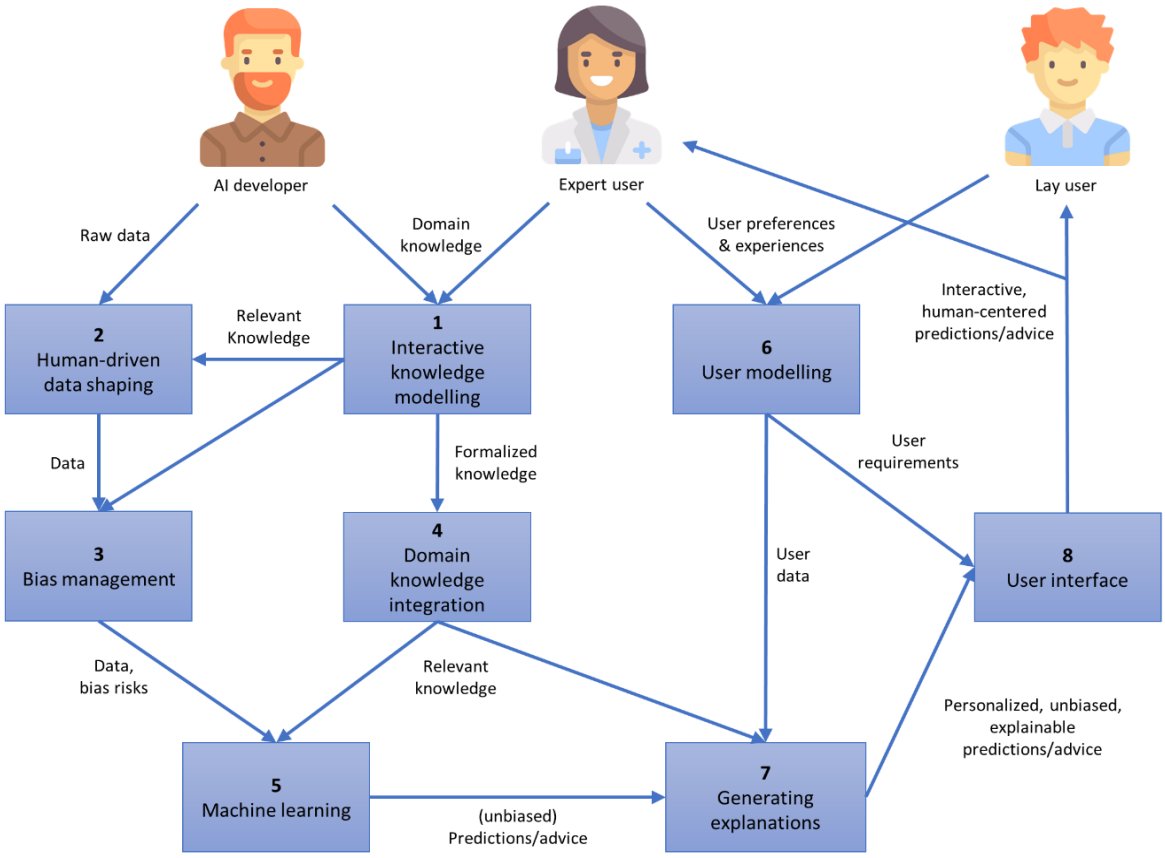

Figure 1: Overview of the integrated FATE system with functional modules.

The individual components that constitute the FATE system are developed in a modular fashion. Typically, these are research topics themselves for which we utilize state-of-the art insights and craft these into a functional module that forms a building block in the integrated FATE system (see Figure 1 for a schematic illustration). The creation of these modules is a combination of implementing the latest scientific insights into a working software module and original research. The functional modules contribute to five research topics that are currently explored within FATE: *I) Fair AI, II) Hybrid AI, III) Explainable AI, IV) User interaction & Co-learning and V) Secure learning*.

I) **Fair AI**, through bias identification, evaluation and mitigation as to ensure a fair and trustworthy system (Module 3, Figure 1). Bias has been touted as the Achilles heel for many AI systems, as it can easily skew predictions into unwanted directions [2]. In some cases you want to avoid bias, for instance when having ML models trained on a particular user set

making predictions for another group of users, while in other cases bias can be exploited, i.e. by making more accurate predictions when the system is aware of existing biases. To ensure a fair and trustworthy AI model, the AI developer takes the following steps to detect and mitigate bias: (1) collect the knowledge for risks, requirements and expectations from experts and stakeholders; (2) with the knowledge, inspect the data for bias risks; (3) based on data and knowledge, priorities in bias risks are selected. The scope, the metrics and the necessary model features are determined; (4) detect bias in impact by running a baseline model and visualizing metrics; (5) act upon the bias risk detection and perform bias mitigation; and finally, (6) evaluate the system by the metrics and use the audit of the process for communication and documentation.

II) **Hybrid AI**, in which we combine knowledge and data-driven approaches to obtain enriched predictions. In Module 1 "Interactive knowledge modelling", domain knowledge is captured in a formal representation by the AI developer and domain expert(s). They keep the knowledge model updated with the latest insights during the system's lifecycle. In Module 2 "Human-driven data shaping" the AI developer transforms the raw data into data that can be used to train the machine learning model. Domain knowledge from Module 1 is used to extract useful and meaningful features. In Module 4 "Domain knowledge integration", domain knowledge is then used to improve predictions that would in a sole data-driven approach perform less well, and/or to provide constraints through which the generated advice can be improved/tailored towards the user-specific context. Domain knowledge can also aid in dealing with bias (topic I) and help with generating explanations of predictions (topic III). Currently, domain knowledge is added to the Machine Learning component (Module 5 in Figure 1) by using a Graph Neural Network [8, 9]. Both knowledge and data are combined into one graph and used to make predictions [21].

III) **Explainable AI**. AI-generated advice – particularly coming from modern approaches such as deep learning models –may be fairly hard to comprehend for (lay) users. However, to obtain an appropriate level of (i.e., calibrated) trust in the system, it is of utmost importance that the system is able to provide adequate explanations that are comprehensible and acceptable to its users [10]. For instance in a healthcare context, both a healthcare professional (e.g., a doctor) consulting an AI-driven prediction system, as well as a patient who may obtain health-related advice from said system need to be able to understand the AI-generated predictions. Not to do so could severely undermine the trust that user place in the system, which can easily result in less effective use. To cater effectively for these different types of user, user modelling is conducted in Module 6 "User modelling". Knowledge about users is captured in two ways: (1) a formal user model containing data of a specific user (e.g., demographics, preferences, interaction history with the system), which is used to personalize explanations, and (2) documentation of user research findings, which is used to inform the user interface design. Based on input from the user module (Module 6), domain knowledge (Module 4) and prediction models (Module 5), explanations are generated in Module 7 "Generating explanations". Information for a variety of explanations (e.g., feature-based, example-based, contrastive) is extracted from the machine learning model. Explanations are personalized based on the user model. An interpretable confidence estimation [7] is also calculated. The generated explanations help users understand the AI model and its predictions. The confidence estimation is particularly useful for trust calibration, because it allows users to adapt their level of trust to the AI's certainty level on a case-by-case basis.

IV) **User interaction & Co-learning**. Users may interact with the system in various capacities and roles. We can identify the following archetypical user roles, although this can vary for different use cases: 1) AI developers/data scientists: persons who train the algorithms and interact with the system from a research/engineering perspective, 2) expert users: persons who interact with the system in their capacity as a professional and as such can provide domain knowledge and expertise that can be incorporated for future iterations, and 3) lay

users: persons who interact with the system from a more lay position and/or receiving role, e.g. a patient. For each of these roles, the manner of interaction and mode of presentation should be tailored, also taking into account a particular context derived from the application domain. Particularly for the expert and lay user roles, it may be important to endow the system with appropriate social capabilities, as this can aid in effective application of ML/AI approaches in a social context [6]. Users interact with the system through a user interface (Module 8)**,** which allows them to interactively explore the AI output through user interface components that are embedded in larger applications (e.g., an AI for medical decision support might be embedded in an Electronic Medical Record). The user interface is designed according to user-centered design principles to ensure that users can use the AI output to reach their goals effectively, efficiently, and enjoyably.

V) **Secure learning**. In many cases, the input data will be sensitive. Various approaches exist to prevent or considerably reduce the risk of information leakage. In the health domain, federated learning is used often for horizontally partitioned data, which is also the case for the use case described in the next section. While this approach provides good protection, it is still possible to extract information from the models that are trained in federated locations. Therefore, we explore how additional approaches, like Multi Party Computation or Differential Privacy, may be used to make the training process more secure.

While the above topics each provide challenges in their own right, it is relevant to integrate these functionalities into one system because there is considerable interplay between them. This may include:

- Explicitly modelled world knowledge is not only used to improve the Hybrid AI component, but will also be needed to more effectively guide the bias detection component, to explicitly model user preferences for the interaction part, and to make better explanations towards the user.
- Bias mitigation measures that are taken by the bias component should be an integral part in the Explainable Artificial Intelligence component: since the user should be warned that either the prediction is inaccurate because of bias, or that this has been accounted for.
- The choice of explanations and the way it is presented to the user, will have impact on the effect. Some types of explanations will not be understood so the system may want to conclude that other forms would be more effective.
- When secure learning methods are used, information is effectively hidden so that only the necessary information is actually known. This puts limits on requirements like transparency and explainability. For instance, explanations using actual examples from the training set may become impossible. Another effect may be that bias identification is not possible when it was not an explicitly formulated information element in a secure learning procedure.

3. A first prototype

To establish sustained progress in the decision-making and task performance for complex problems, hybrid AI-solutions are being proposed and researched that complement and reinforce human's cognitive work [11,12]. We developed a first prototype of the FATE hybrid system that incorporates basic functionalities for bias management, personalized explanations, knowledge modelling and federated learning in order to establish ongoing, effective and fair, *human-AI co-learning*.

The research and development of such hybrid intelligence is *transdisciplinary* by nature (cf, [13]), requiring transdisciplinary system modelling (cf. [14]). The FATE project team comprises the diverse disciplines for researching the relevant topics with their dependencies (such as human factors, agent modelling, machine learning, human-computer interaction, behaviour change, chronic diseases). For establishing the joint project objectives, the research questions and the research & development methods, we are developing an easy-to-understand modelling language with a common vocabulary that describes the envisioned co-learning solution as re-usable design patterns (cf., [15,16,17]). In addition to this design approach, for the evaluation we specified and assessed the first set of evaluation methods and metrics, focusing on understandability and trust of explained AI-output [18].

We used the socio-cognitive engineering method [5] to facilitate a common ground between the team members from different disciplines, and to integrate the relevant technical, human factors, and operational aspects into a coherent problem description and solution direction.

The design of the scenario and corresponding prototype forced the project team to harmonize the machine learning and knowledge modelling, to collaborate on the transdisciplinary research topics of fairness and explainable AI, and to integrate AI's input and output into a coherent set of interaction design patterns (cf. [19]). For example, explanations can help to become aware of bias risks and to mitigate biases, with the specific roles of AI-developer, domain expert (Health Care Professional, HCP) and layperson (patient) in a diabetes use case.

Docker[1] was used to integrate the different modules from Figure 1 into a single FATE prototype. The user interface module was built using Dash[2] and Bootstrap[3] and contains a set of UI components that can be combined in a way that is tailored to the user interacting with the system (e.g., a HCP may require different components than a patient). Figure 2 shows a screenshot of the prototype, including UI components for the certainty of an AI prediction, evidence for and against a prediction, a contrastive explanation, an example-based explanation, and the performance of the AI model for a specific prediction.

[1] Docker, Inc (n.d.). *Docker: Empowering App Development for Developers*. https://www.docker.com/
[2] Plotly (n.d.). *Introduction to Dash*. https://dash.plotly.com/introduction
[3] *Bootstrap – The most popular HTM, CSS, and JS library in the world.* (n.d.). https://getbootstrap.com/

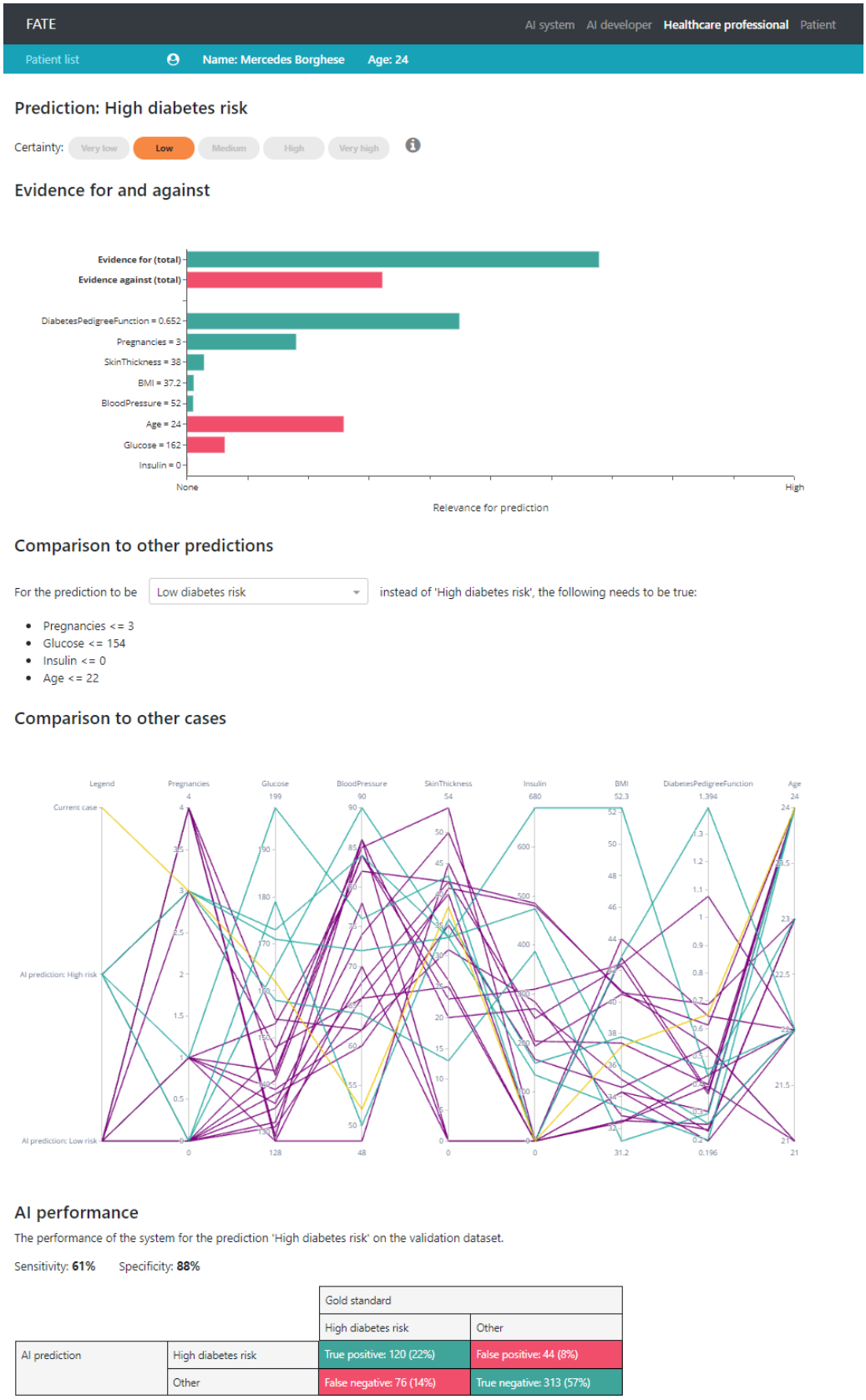

Figure 2. Screenshot of the FATE prototype for the HCP user role. This example shows UI components for the certainty of an AI prediction, evidence for and against a prediction, a contrastive explanation, an example-based explanation, and the performance of the AI model for a specific prediction (from top to bottom).

4. Outlook

The FATE project is a multi-year research project that is continuing in 2021. Compared to last year, there will be a stronger focus on the user-roles and the interaction that they will have with the system. In particular, we will focus more on the research topics of of fairness, explainable AI, co-learning, and secure learning. We continue the goal of building a generic system, meaning that it needs to be able to deal with a wide variety of different contexts. Context can change, depending on both different user roles and different use cases. The three user roles remain the same, but we generalize these more so that they can be applied within use cases in (very) different domains:

1) Researcher (a domain researcher, e.g. a medical researcher, a criminologist, a sociologist),
2) Consultant (advisor of a subject, e.g. a medical doctor, a judge, a civil servant),
3) Subject (affected person, e.g., a patient/consumer, a convicted person, a naïve users, etc.).

Each of these user roles will place specific demands on the system with respect to the type of interaction that will occur over time.

The system will be adaptive towards the users, learn from the interaction and the feedback that users provide and from additional data that may become available as scenarios unfold, and thus evolve towards an increasingly personalized and optimized support system that can provide advice in a fair and secure manner.

Acknowledgements

The FATE project is funded by the TNO Appl.AI program (internal AI program).

References

[1] Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. (2019, June). Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning* (pp. 1-4).

[2] Zou J, Schiebinger L. AI can be sexist and racist - it's time to make it fair. Nature. 2018 Jul;559(7714):324-326. doi: 10.1038/d41586-018-05707-8. PMID: 30018439.

[3] Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. *arXiv preprint arXiv:1901.00770*.

[4] Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, *25*(1-2), 161-197.

[5] Neerincx, M., Vught, W., Blanson Henkemans, O., Oleari, E., Broekens, J., Peters, R., ... & Bierman, B. (2019). Socio-Cognitive Engineering of a Robotic Partner for Child's Diabetes Self-Management. Frontiers in Robotics and AI, 6, 118-134. https://doi.org/10.3389/frobt.2019.00118

[6] de Greeff, Joachim, and Tony Belpaeme. "Why robots should be social: Enhancing machine learning through social human-robot interaction." *PLoS one* 10.9 (2015): e0138061.

[7] van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *Int. J. of Human-Computer Studies*, 102493.

[8] Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., ... & Sun, M. (2018). Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*.

[9] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

[10] De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, *12*(2), 459-478.

[11] Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., ... & Hung, H. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, *53*(8), 18-28.
[12] Peeters, M. M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2020). Hybrid collective intelligence in a human–AI society. *AI & SOCIETY*, 1-22.
[13] Lang, D. J., Wiek, A., Bergmann, M., Stauffacher, M., Martens, P., Moll, P., ... & Thomas, C. J. (2012). Transdisciplinary research in sustainability science: practice, principles, and challenges. *Sustainability science*, *7*(1), 25-43.
[14] Karaca, Y., & Altuntaş, E. Y. (2020, July). Decision Tree-Based Transdisciplinary Systems Modelling for Cognitive Status in Neurological Diseases. In *International Conference on Computational Science and Its Applications* (pp. 442-457). Springer, Cham.
[15] Van Diggelen, J., Neerincx, M., Peeters, M., & Schraagen, J. M. (2018). Developing effective and resilient human-agent teamwork using team design patterns. *IEEE int. systems*, *34*(2), 15-24.
[16] van der Waa, J., van Diggelen, J., Siebert, L. C., Neerincx, M., & Jonker, C. (2020). Allocation of Moral Decision-Making in Human-Agent Teams: A Pattern Approach. In *International Conference on Human-Computer Interaction* (pp. 203-220). Springer, Cham.
[17] van Harmelen, F., & ten Teije, A. (2019). A Boxology of Design Patterns for Hybrid Learning and Reasoning Systems. *Journal of Web Engineering*.
[18] van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2020). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 103404.
[19] Neerincx, M. A., van der Waa, J., Kaptein, F., & van Diggelen, J. (2018). Using perceptual and cognitive explanations for enhanced human-agent team performance. In *International Conference on Engineering Psychology and Cognitive Ergonomics* (pp. 204-214). Springer, Cham.
[20] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.
[21] de Boer, M. & Hillerström, F. (2021). Hybrid AI using Graph Neural Networks. In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.