



Human-centric algebraic machine learning

Human-AML Interaction (Initial)

Deliverable D4.1

WP4 - Human AML Interaction



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952091.

Consortium



ALGEBRAIC AI



Chantalimaud
Foundation



German
Research Center
for Artificial
Intelligence



TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

uc3m



Project

Title: ALMA: Human Centric Algebraic Machine Learning
 Acronym: ALMA
 Coordinator: Proyectos y Sistemas de Mantenimiento SL (eProsima)
 Grant agreement: 952091
 Call: H2020-EIC-FETPROACT-2019
 Program: Horizon 2020
 Start: 01 September 2020
 Duration: 48 months
 Website: <https://alma-ai.eu>
 E-mail: alma@eprosima.com
 Consortium: **Proyectos y Sistemas de Mantenimiento SL (eProsima) (EPROS)**, Spain
Champalimaud Foundation (CF), Portugal
ALGEBRAIC AI (ALG), Spain
German Research Center for Artificial Intelligence (DFKI), Germany
Inria (Inria), France
TU Kaiserslautern (TUK), Germany
FIWARE Foundation e.V. (FF), Germany
Technical Research Centre of Finland (VTT), Finland
Carlos III University (UC3M), Spain

Deliverable

Number: **D4.1**
 Title: **Human-AML Interaction (Initial)**
 Month: 18
 Work Package: WP4 - Human AML Interaction
 Work Package leader: Inria
 Deliverable leader: National Institute for Research in Digital Science and Technology (Inria)
 Deliverable type: R: Document, report
 Dissemination level: Public (PU)
 Date of submission: 2022-03-16
 Version: v1.0
 Status: Not finished

Version history

Version	Date	Responsible	Author/Reviewer	Comments
v1.0	2022-02-28	Inria	Janin Koch	Partial Draft
v1.1	2022-03-13	Inria	Janin Koch, Vitor Fortes Rey	Partial Revision
v1.1	2022-03-15	Inria	Janin Koch, Wendy Mackay, Vitor Fortes Rey	Partial Revision
v1.2	2022-03-17	Inria	Janin Koch, Wendy Mackay, Vitor Fortes Rey	Final Revision

Executive summary

In this document, D4.1 - Human-AML Interaction (Initial), we present an overview of the literature on mutual understanding in collaboration. The objective of this deliverable is to extend knowledge on human-AI understanding in order to lay an foundation for designing and implementing effective human-computer interactions in which users remain in control.

This deliverable D4.1 - Human-AML Interaction (Initial) gives a brief introduction of what collaboration is and how it relates to understanding and explainability. We provide insights from both a human-centric and an AI-centric perspective on developing a common understanding between humans and machines, as well as the aspect of explainability within this interaction. We explain how social science studies on grounding can support researchers in their investigation of human-AI interactions. Current research on human-centered explainable AI (HCXAI) and the challenges it presents for explainable AI (xAI) system design are discussed. Finally, we present existing xAI research including the limitations we identified in some of these approaches for achieving successful human-AI collaboration.

This deliverable D4.1 - Human-AML Interaction (Initial) is the foundation for our work on human-AI interaction methodology, D4.4 - Interaction Paradigm Methodology (Initial) in which we reported on our exploratory research into new interaction methodologies for enabling more user-centered human-AML interaction. It also influenced our work on exploring AML-explainability capacities for creative application in practice, D4.7 - Working Prototype and Designs (Initial).

This deliverable will be followed by D4.2 Human-AML Interaction (Revised), in which we will extend this review and analysis of the literature to include other aspects of human-AI interaction. We will focus primarily on cognitive aspects of *agency*, as well as outlining how to support *theory of mind* for enabling more sophisticated human-AI interaction.

Contents

Executive Summary	3
Contents	4
1 Introduction	6
2 Grounding in Human-AI Understanding	6
2.1 Grounding in Human-Human Conversations	7
2.2 Supporting Grounding in a Human-AI Collaboration	7
3 Human-centered Explainable AI (HCXAI)	8
3.1 Socio-technical approaches to human-centered explainable AI	9
3.2 Challenges and opportunities of human-centered explainable AI	9
3.3 Evaluation methods for human-centered explainable AI	9
4 Explainable AI (XAI)	10
4.1 Model-specific	10
4.2 Explanation by simplification	11
4.3 Visual explanation	11
4.4 Local and Global Explanation	11
5 Discussion	11
6 Conclusions	12
References	13

1 Introduction

The goal of this deliverable is to give insight into current research of the role that human and machines play in human-AI interaction. While there are a number of factors involved in the concept of collaborating on a more equal level with a system, this initial deliverable D4.1 focuses on the the aspect of mutual understanding within collaboration. This deliverable outlines relevant developments in these fields using state-of-the-art taxonomies and frameworks.

Before delving into relevant literature on understanding and explainability, we first highlight some definitions of collaboration that serve as the foundation for this deliverable's understanding of human-AI collaboration. The key phenomenon in collaboration is referred to as “intersubjectivity” in the social sciences and philosophy. Intersubjectivity refers to how two or more minds inter-relate: how they understand and work together from their individual cognitive positions [1]. Some well-known social theories related to intersubjectivity are *mediated cognition* [2], *situated learning* [3], *knowledge building* [4], and *distributed cognition* [5]; Wood et al. present an overview of differences among these perspectives [6].

In this context, collaboration is defined as “a process through which parties who see different aspects of a problem can constructively explore their differences and search for solutions that go beyond their own limited vision of what is possible” [7].

In a fourth definition, Roschelle et al. define collaboration as “a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem” [8]. This definition originated in the field of collaborative learning and is probably closest to the understanding of collaboration in human computer interaction (HCI) research. It builds on the notion of collaboration as a cognitive action but also includes aspects of the previously mentioned definitions. Some empirical evidence exists that such collaborative learning enhances the cognitive capabilities of the people involved, allowing them as a team to reach a level of cognitive performance that exceeds the sum of the individuals’ [9]. This is comparable to insights from creative collaborations, where interactions between contributors became more important to the creative outcome than the creative capacity of each individual [10]. The group’s creative capabilities therefore often exceed the sum of individual skills, to produce ideas that are both novel and of high quality [11, 12], a phenomenon often referred to as “collaborative creativity”.

This collaborative potential has attracted the interest of the HCI and AI research communities, especially in the area of advancing human-AI collaborative approaches. These collaborative effects, however, are based on interaction and common understanding, as described by Rochelle et al. as “continued attempt to construct and maintain a shared conception of a problem” [8]. As a result, in order to investigate these aspects, this deliverable focuses on insights about forming a common understanding between humans and machines, as well as the aspect of explainability, from both a human-centric and a AI-centric perspective.

This deliverable describe in the remain of this document current research from three different perspectives: Conversational grounding for human-AI interactions; Human-centric perspective on explainability; and AI-centric perspective on explainability (xAI). Combining these three perspectives allows us to create a space of possibilities to explore human-AI explainability in our related deliverables.

2 Grounding in Human-AI Understanding

“*Grounding*” is the ability to create shared bases of knowledge, beliefs, and assumptions about a common goal [13]. While arguably not a complete model of collaborative behaviours, multiple studies demonstrated its explanatory power in small-scale, dyadic collaborations [14].

2.1 Grounding in Human-Human Conversations

We use this term here in the sense employed by Clark et al. from the tradition of conversational analysis [15]. In this seminal work, they argue that establishing common ground is the foundation of a number of joint activities, including collaboration and communication.

This concept has informed a number of theoretical frameworks, including in AI, and in particular in the context of dyadic interactions, e.g. the collaborative model [16], the Mixed-Initiative (MI) framework [17], and theories of collaborative learning [13]. *Grounding* emphasizes the need for efficient communication to establish a collective understanding of the shared goal, by ensuring that every contribution to the collaboration is both clearly expressed and correctly received by its addressees.

Clark et al. argue that successful grounding is influenced by two factors: purpose and medium [15]. *Purpose* is the objective, desire, or emotion that should be conveyed in a collaborative endeavor. The *medium* is the means used to express one's current purpose as well as its costs. They also introduce the notion of 'least collaborative effort', which is the phenomenon that collaborators will often try to communicate as little as possible – yet as much as needed – and with the most effective medium to enable successful reception.

2.2 Supporting Grounding in a Human-AI Collaboration

To this end, Koch et al. identified four key requirements for supporting grounding in a human-AI collaboration [18]:

- 1) *Expressing one's own objectives*: Grounding relies on the successful expression of the collaborators' objectives, requirements, and intents. This is not a trivial endeavor when collaborating with a computer. Depending on the communication channel, an AI system needs to formulate its concepts in ways and formats that can then be presented to other collaborators (e.g., a suggestion into sufficiently descriptive words). Existing examples include dialogue systems applying Belief–Desire–Intention models [19] and theoretical models that can generate purposeful speech acts [20] to express meaningful objectives. There is also a growing body of research exploring the opportunities of concept learning [21, 22], which aims to enable systems to combine goals and communicate them in association with existing concepts.
- 2) *Selecting the most effective medium*: Collaborators need to select the medium(s) that best convey their purposes. In human-to-human conversation, a purpose can be expressed in any combination of verbal and non-verbal channels – e.g., speech, pointing, body language, and even extended media such as writing or drawing. The choice of medium depends on the availability of these channels, on the effort and skill they require, and on the expected ability of the addressee to successfully perceive and interpret that purpose. A computer's choice of mediums, and its ability to exploit them, is necessarily limited by its physical capabilities. Robots that share the same space and media as their human collaborators can engage in three-dimensional pointing, object manipulation, and even eye movement [23]. Virtual agents, on the other hand, are limited in their capacity to refer to physical elements but might engage in visual, animated, or written representations of information as comparatively strong expressive mediums.
- 3) *Evaluating the effort of an action*: Clark and Wilkes-Gibbs introduced the principle of "least collaborative effort" [16] as a trade-off between the initial effort of formulating and expressing a message, and the effort of rephrasing that initial message when it is not properly understood. Previous work on that principle has mainly focused on creating short and precise communication efforts for which accurate wording was identified as an optimal strategy, commonly seen in conversational bots. In contrast, Clark et al. showed that the least *collaborative* effort in joint work can follow different patterns and relaxed constraints. The effort of understanding is shared within a group, and groups can accept "wordings" with internal references that need not be unique to the given context. This relaxation presents both opportunities and challenges for AI systems. The conversation structure of most conversational agents, such as

Siri [24], follows that least effort principle by providing short and specific answers. Extending this to a least-*collaborative*-effort strategy would require to connect knowledge with previous, contextual, and general expressions. N. Mavridis introduced the notion of ‘situated language’ to overcome these issues and to extend a computer’s communication capabilities to time- and location-independent references [23].

4) *Confirming reception of the initial objective*: Successful conversational grounding also requires collaborators to find evidence that their meaning or objective was successfully received and interpreted by other members [15]. Such evidence may be positive, when the message was successfully understood, or negative when not. Human-to-human positive evidence may take the form of positive acknowledgement, such as nodding or interjecting ‘mmm-hmm’, or be demonstrated by relevant next-turn response. Naturally, the selected medium might affect the reaction to the received expression. Enabling a machine to assess successful reception evidence from other group members requires new research into not just natural-language processing in relation to natural interaction [25] but also into the handling of non-verbal behaviour [26].

The four key requirements for supporting grounding in a human-AI collaboration outlined above have the potential to inform human-AI collaboration frameworks regarding how reciprocal understanding among participants can be achieved.

3 Human-centered Explainable AI (HCXAI)

While the letters “HCI” do not exist in “XAI”, explainability in AI is just as much, if not more, a Human-Computer Interaction (HCI) concern. However, in XAI’s technical discourse, the human aspect of the equation is frequently overlooked [27]. There is a number of research fields at the intersection of human-computer interaction (HCI) and AI often referred to as Human-centered explainable AI (HCXAI) [27], human-centered artificial intelligence (HCAI) [28] or also called human-centered machine learning (HCML) [29] that emerged over the last decade to address these issues.

The focus in these research fields is on how humans can better understand intelligent systems’ contributions, explanations, and, in part, how humans can better explain themselves to the system to enable a better and more effective interaction between a human and AI. To what extent this more effective interaction aids to improve a *tool use*, e.g. as in xAI; a *cooperation*, e.g. in mixed-initiative interactions systems; or a *human-computer collaboration*, e.g. in human-computer partnership systems; is not clearly distinguished within these fields. In this section, we mainly refer to HCXAI because it is the only field solely focused on explainability; however, many similar approaches can be found in the other research areas as well.

When it comes to explaining a system’s contribution, the difference between these fields and general explainable AI (XAI) research is in the context of the explanation: xAI refers to *how* the system works and makes decisions, whereas HCXAI seeks to fulfill a need for explanations to *why* a given result is relevant to the current task. The *human-centered artificial intelligence* (HCAI) framework [28] clearly demonstrates this shift in emphasis when compared to xAI (see section 4). The key factors for a more human-centred AI are:

- Design for high levels of human control as well as high levels of computer automation in order to improve human performance.
- Recognize the situations in which full human control or full computer control is required.
- Avoid the dangers of having too much human control or too much computer control.

In their overview of the human-centered artificial intelligence framework, Shneiderman et al. describe how these elements are required to develop more reliable and trustworthy human-AI relationships, emphasizing the importance of explainability and controllability in this context [28].

3.1 Socio-technical approaches to human-centered explainable AI

This requires that HCXAI system designers have a more holistic comprehension of the situation and users of their systems. This inspired a number of researchers to advocate for a more socio-technical approach to human-AI explainability. A socio-technical approach to HCXAI, according to Ehsan et al. [27], requires a more holistic understanding of the socially situated characteristics of XAI systems, as well as taking social and technological elements into account when building such systems.

To understand the "who" in XAI and communicate their needs and goals, XAI developers must first place the "who" in the sociotechnical context [30]. This could be accomplished by actively including the community in the design processes of explainable AI. Zhu et al. present a full review of explainable AI for designers and co-creative practices, especially in the domain of game design [31], and even suggests to divide HCXAI research into subcategories concentrating on a specific "who," in this example explainable AI for designers (XAID).

Simultaneously, Liao et al. emphasize that for XAI users, an actionable understanding is often a socially situated understanding, allowing them to comprehend not just the technology component but also the socio-technical system as a whole [30]. Designing for such situated understanding requires xAI developers gaining hands-on experience with the social situation in which the xAI will be used, which can be challenging.

3.2 Challenges and opportunities of human-centered explainable AI

There are two main challenges with HCXAI using a socio-technical approach as already hinted in the paragraph above. First it requires constructive engagement with partner communities of practice, to identify the needs and norms of the targeted community [27]. Participatory Design is one approach that has been highlighted in this context because it reduces the power dynamics between the designer and the user and aims to support democratic values at every stage of the design process [27]. This approach fits in nicely with the key properties of HCXAI and can help bring underrepresented viewpoints to the spotlight as it can include a number of perspectives.

The second main challenge is the requirements of researchers that are familiar with the social situation in which the xAI will be used. This would require active translational efforts on the part of a wide group of practitioners and scholars. This means that, in comparison to researchers with intellectual depth in one field, we need more researchers with intellectual depth in two (or more) areas and the capacity to bridge these domains [27]. A number of initiatives have already started to cross disciplines, e.g. [32], to create a joint effort for promoting more relevant human-centered explainable AI.

3.3 Evaluation methods for human-centered explainable AI

Evaluating human-centered explainable AI is challenging, and numerous explainability methods and theoretical evaluation scores have been proposed [33]. Nevertheless, an increasing body of research work started investigating these evaluation methods of explainable AI from a human-centered perspective for their usefulness and applicability [33, 29, 34, 35, 36, 37].

For example, Fel et al. studied how useful human-centered explainability approaches are in real-world contexts and how well theoretical metrics predict their relevance for human use [33]. They evaluated the usefulness of six representative attribution approaches for binary categorization tasks using a series of large-scale online psychophysics experiments. The study found that the degree to which attribution approaches were useful to human observers varied greatly among categorization tasks, and that the trustworthiness and usefulness measures appeared to have only a weak correlation [33]. While explainable

AI is often referenced to increase trustworthiness, Fel et al. challenge this merge and ask for more research in how to untangle evaluation methods for trust and usefulness of explainabilities [33].

Similarly, Sperrle et al. emphasize the complexity and interconnection of measures in a large review of human-centered evaluations (HCE) of human-centered machine learning (HCML) [29]. They also present dimensions derived from the findings of a comprehensive survey to structure evaluation methods of such systems, including experiment setups, expertise of participants to common data types collected and task analysis applied. The authors further give a checklist as guidance for the design of HCML evaluations and a template for reporting HCE results to help with the evaluation of human-centered explainable AI. [29] Notably, similar to some of Fel et al.'s observations [33], the characteristics or properties explained to users were model quality, perceived quality, model transparency, trustworthiness, as well as interpretability and controllability [29].

Work package four focuses on extending the explainability and controllability of human-AML interaction based on the interpretability of the models. According to the survey, these are still under explored research fields, with only 14% of papers assessing model interpretability and only 20% of studies aiming to give controllability for users, and only two studies evaluating if controllability was achieved [29]. These findings serve as the foundation for our investigation of the intractability, explainability, and possibility for human control of AML models.

4 Explainable AI (XAI)

The goal of an explainable AI system is to provide explanations that make its behavior more understandable to humans. The following recommendations are intended to improve the design of these systems [38]. An XAI system should:

- describe its capabilities and understandings within the current context;
- explain what it has done, what it is doing now, and what will happen next;
- disclose the underlying information it is acting upon [39]

The field of XAI is a very active area of research full of specialized (and often re-defined) terms. For more details, the ontology and taxonomy of XAI at a high level can be found in [40]. Following the taxonomy from [40], Here we will go briefly over some of its concepts and how they relate to mutual understanding and collaboration.

4.1 Model-specific

Model-specific XAI approaches are those that use a specific model structure to bring transparency to its decisions [40]. For example, in a linear model or small decision tree, the inherent model structure can be used to provide explanations.

Still, as pointed out by [30], the effectiveness of explanation is not intrinsic to the model but lies in the perception and reception of the person receiving the explanation. Even when presented with a completely transparent model there is no guarantee that the person at the receiving end can make sense of all the information, or is not overwhelmed. The quality of an explanation lies in its ability to provide appropriate information that can be understood and utilized [30]. This is contingent on the receiver's current knowledge and their goal for receiving the explanation, among other human factors and can only be achieved when grounding is achieved, that is, when reciprocal understanding between the AI and participant's goals is reached as explained in Section 2.2.

4.2 Explanation by simplification

Explanation by simplification consists of approximating a black box model with one that is more transparent and interpretable. If the simplified model is not directly used in the decision-making process, but only used to provide explanations, then one must be aware of the gap between explanations as provided by the simplified model. If its output does not align to the original complex one, a low fidelity simplified model should not engender trust in its explanation or even in the original complex model [41]. Thus, one has to be careful in the grounding process to provide appropriate information so that the conversational grounding is valid.

4.3 Visual explanation

This type of XAI is based on visualization techniques to explain decisions. As described in Section 2.2, it is important to select the most effective medium to convey the collaborator's purposes. One may be tempted to use the latest visualization technique, but this remains challenging for complex models. For example, in the case of image classification with convolutional neural networks, creating a heatmap showing which pixels contributed more to the image's classification. Those are popular for deep learning methods in general, but, even if useful for debugging, they can be misleading in human-AI communication as they tend to highlight edges and may even show similar heatmaps when providing explanations for different classes [41].

4.4 Local and Global Explanation

Local explanations are those that approximate the model around a specific example of interest, providing information on how the model behaves in this narrow area. This is opposed to global explanations, where the overall logic model-wide is explained, that is, a global-view is provided [30]. For example, a local explanation for a decision tree would display only the decision path followed by that example and points close to it, while a global one would include all model logic. It is clear that, given an overall transparent model, one can provide more control when global explanations are also provided. This is even more important as the number of features increases and points may be sparse due to the curse of dimensionality.

That is not to say local explanations are not useful, they can quantify the importance of features as is the case in explanation by feature relevance, which is similar to simplification, but focuses on providing transparency by evaluating features based on their expected marginal contribution to the prediction being explained. Examples include LIME [36] and SHAP [42]. Those can provide insights into current predictions, but alone cannot guarantee effective explanations [43].

5 Discussion

This literature inspired our approach to explore a more explainable human-AI interaction that allows the end-user to better understand and control the AML-system.

The four key requirements for supporting grounding in a human-AI collaboration outlined earlier have the potential to inform human-AI collaboration frameworks regarding how reciprocal understanding among participants can be achieved, see section 2. Thinking about the explainability objective, the medium we use, and the user's interpretation of reception pushed our ideas for human-AML interaction in D4.7 - Working Prototype and Designs (Initial) toward a more targeted mapping of specific movements, locomotion — font, and expressiveness — color style, while keeping all aspects coherent in a single explainability visualization.

The inclusion of diverse end-users in the process of building successful human-centered explainable AI

is highlighted by related work, see section 3. To address this limitation, we conducted a participatory design workshop with a wide variety of end-users with diverse needs, on which we report in deliverable D4.4 - Interaction Paradigm Methodology (Initial).

We have identified many challenges, which led us to study how to provide effective model-specific explanations for the outputs of AML models. While still under development, our current focus is on building simplified but transparent models and, through participatory design, develop valid visual explanations that are informative and correct from a grounding as well as machine learning perspective.

6 Conclusions

The goal of this deliverable D4.1 - Human-AML Interaction (Initial) is to give insight into current research on the the aspect of mutual understanding within human-AI collaboration. This addressed the role that human and machines can play in human-AI interaction as well as their reciprocal understanding of objectives, intends and expressions.

We presented insights from both a human-centric and an AI-centric perspective on developing a common understanding between humans and machines, as well as the aspect of explainability within this interaction. We describe how social science research on grounding can help researchers explore human-AI interactions. We continue to discuss current research on human-centered explainable AI, the challenges it presents for system design, and the opportunity for participatory design in this context. Finally, we present current research in system-centric explainable AI based on a new taxonomy of xAI techniques, as well as some of the limitations that we have identified with some of these approaches. We conclude with a discussion how this literature impacted our approaches which we report in deliverable D4.4 - Interaction Paradigm Methodology (Initial) and D4.7 - Working Prototype and Designs (Initial).

This deliverable will be followed by D4.2 Human-AML Interaction (Revised), in which we will extend this review and analysis of the literature to include other aspects of human-AI interaction. We will focus primarily on cognitive aspects of *agency*, as well as outlining how to support *theory of mind* for enabling more sophisticated human-AI interaction.

References

- [1] Gerry Stahl. “From intersubjectivity to group cognition”. In: *Computer Supported Cooperative Work (CSCW)* 25.4-5 (2016), pp. 355–384.
- [2] Phil Turner. *Mediated Cognition*. Cham: Springer International Publishing, 2016, pp. 27–40. ISBN: 978-3-319-42235-0.
- [3] Jean Lave and Etienne Wenger. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.
- [4] Huang-Yao Hong et al. “Teacher-education students’ views about knowledge building theory and practice”. In: *Instructional Science* 39.4 (2011), pp. 467–482.
- [5] James Hollan, Edwin Hutchins, and David Kirsh. “Distributed cognition: toward a new foundation for human-computer interaction research”. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 7.2 (2000), pp. 174–196.
- [6] Donna J Wood and Barbara Gray. “Toward a comprehensive theory of collaboration”. In: *The Journal of Applied Behavioral Science* 27.2 (1991), pp. 139–162.
- [7] Barbara Gray. “Collaborating: Finding common ground for multiparty problems”. In: (1989).
- [8] Jeremy Roschelle, Stephanie D Teasley, et al. “The construction of shared knowledge in collaborative problem solving”. In: *Computer-supported collaborative learning*. Vol. 128. 1995, pp. 69–197.
- [9] Michael J Baker. “Collaboration in collaborative learning”. In: *Interaction Studies* 16.3 (2015), pp. 451–473.
- [10] R Keith Sawyer and Stacy DeZutter. “Distributed creativity: How collective creations emerge from collaboration.” In: *Psychology of aesthetics, creativity, and the arts* 3.2 (2009), p. 81.
- [11] Cecilia R Aragon and Alison Williams. “Collaborative creativity: a complex systems model with distributed affect”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2011, pp. 1875–1884.
- [12] Sonja Gutwein. “Computer support for collaborative creativity”. In: *Beyond the Desktop* (2013), p. 15.
- [13] Michael Baker et al. “The role of grounding in collaborative learning tasks”. In: *Collaborative learning: Cognitive and computational approaches* 31 (1999), p. 63.
- [14] G Stahl. “Shared meaning, common ground, group cognition”. In: *Group cognition: Computer support for building collaborative knowledge* (2006), pp. 347–360.
- [15] Herbert H Clark, Susan E Brennan, et al. “Grounding in communication”. In: *Perspectives on socially shared cognition* 13.1991 (1991), pp. 127–149.
- [16] Herbert H Clark and Deanna Wilkes-Gibbs. “Referring as a collaborative process”. In: *Cognition* 22.1 (1986), pp. 1–39.
- [17] JE Allen, Curry I Guinn, and E Horvitz. “Mixed-initiative interaction”. In: *IEEE Intelligent Systems and their Applications* 14.5 (1999), pp. 14–23.
- [18] Janin Koch and Antti Oulasvirta. “Group cognition and collaborative ai”. In: *Human and Machine Learning*. Springer, 2018, pp. 293–312.
- [19] Daniel Jurafsky and James H Martin. “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”. In: (2014).
- [20] Philip R Cohen and C Raymond Perrault. “Elements of a plan-based theory of speech acts”. In: *Cognitive science* 3.3 (1979), pp. 177–212.
- [21] Kenneth A De Jong, William M Spears, and Diana F Gordon. “Using genetic algorithms for concept learning”. In: *Machine learning* 13.2-3 (1993), pp. 161–188.
- [22] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. “Human-level concept learning through probabilistic program induction”. In: *Science* 350.6266 (2015), pp. 1332–1338.
- [23] Nikolaos Mavridis. “A review of verbal and non-verbal human-robot interactive communication”. In: *Robotics and Autonomous Systems* 63 (2015), pp. 22–35.
- [24] Apple iOS Siri. *Apple*. 2013.

- [25] Erik Cambria and Bebo White. “Jumping NLP curves: A review of natural language processing research”. In: *IEEE Computational intelligence magazine* 9.2 (2014), pp. 48–57.
- [26] Rana El Kaliouby and Peter Robinson. “Real-time inference of complex mental states from facial expressions and head gestures”. In: *Real-time vision for human-computer interaction*. Springer, 2005, pp. 181–200.
- [27] Upol Ehsan and Mark O Riedl. “Human-centered explainable ai: towards a reflective sociotechnical approach”. In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 449–466.
- [28] Ben Shneiderman. “Human-centered artificial intelligence: Reliable, safe & trustworthy”. In: *International Journal of Human-Computer Interaction* 36.6 (2020), pp. 495–504.
- [29] Fabian Sperrle et al. “A Survey of Human-Centered Evaluations in Human-Centered Machine Learning”. In: *Computer Graphics Forum*. Vol. 40. 3. Wiley Online Library. 2021, pp. 543–568.
- [30] Q Vera Liao and Kush R Varshney. “Human-Centered Explainable AI (XAI): From Algorithms to User Experiences”. In: *arXiv preprint arXiv:2110.10790* (2021).
- [31] Jichen Zhu et al. “Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation”. In: *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE. 2018, pp. 1–8.
- [32] *Human-Centered Explainable AI (HCXAI) workshop at CHI 2022*. URL: <https://hcxai.jimdosite.com>.
- [33] Thomas Fel et al. “What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods”. In: *arXiv preprint arXiv:2112.04417* (2021).
- [34] Peter Hase and Mohit Bansal. “Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?”. In: *arXiv preprint arXiv:2005.01831* (2020).
- [35] Dong Nguyen. “Comparing automatic and human evaluation of local explanations for text classification”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 1069–1078.
- [36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ““ Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.
- [37] Hua Shen and Ting-Hao Huang. “How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 8. 2020, pp. 168–172.
- [38] David Gunning et al. “XAI—Explainable artificial intelligence”. In: *Science Robotics* 4.37 (2019), eaay7120.
- [39] Victoria Bellotti and Keith Edwards. “Intelligibility and accountability: human considerations in context-aware systems”. In: *Human-Computer Interaction* 16.2-4 (2001), pp. 193–212.
- [40] Plamen P Angelov et al. “Explainable artificial intelligence: an analytical review”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11.5 (2021), e1424.
- [41] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.
- [42] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).
- [43] Dylan Slack et al. “Fooling lime and shap: Adversarial attacks on post hoc explanation methods”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 180–186.