Technology in Society 79 (2024) 102685

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Explainable AI for all - A roadmap for inclusive XAI for people with cognitive disabilities

Myrthe L. Tielman [a,*], Mari Carmen Suárez-Figueroa [b], Arne Jönsson [c], Mark A. Neerincx [a,d], Luciano Cavalcante Siebert [a]

[a] Delft University of Technology, van Mourik Broekmanweg 6, Delft, 2826 XE, Netherlands
[b] Ontology Engineering Group (OEG) and Universidad Politécnica de Madrid (UPM), Campus de Montegancedo sn., Boadilla del Monte, Madrid, 28660, Spain
[c] Department of Computer and Information Science, Linköping University, Linköping, 581 83, Sweden
[d] TNO Human Factors, Kampweg 55, Soesterberg, 3769, DE, Netherlands

ARTICLE INFO

Keywords:
Explainable AI (XAI)
Cognitive disability
Responsible AI

ABSTRACT

Artificial intelligence (AI) is increasingly prevalent in our daily lives, setting specific requirements for responsible development and deployment: The AI should be explainable and inclusive. Despite substantial research and development investment in explainable AI, there is a lack of effort into making AI explainable and inclusive to people with cognitive disabilities as well. In this paper, we present the first steps towards this research topic. We argue that three main questions guide this research, namely: 1) How explainable should a system be?; 2) What level of understanding can the user reach, and what is the right type of explanation to help them reach this level?; and 3) How can we implement an AI system that can generate the necessary explanations? We present the current state of the art in research on these three topics, the current open questions and the next steps. Finally, we present the challenges specific to bringing these three research topics together, in order to eventually be able to answer the question of how to make AI systems explainable also to people with cognitive disabilities.

1. Introduction

Artificial Intelligence (AI), in combination with other Information and Communication Technologies (ICT) like smart sensors, cloud computing, conversational agents and virtual reality, is becoming part of our daily lives. AI technology is affecting people in a variety of processes and activities, both professionally and personally. Some examples are solutions for job application processing [1,2], investigation of criminal activities [3,4], e-learning services [5] and healthcare applications [6,7,8]. Acknowledging the opportunities and risks of AI-deployment, research networks (e.g., CLAIRE, TAILOR) and research programmes (e.g., HUMANE AI) have started to investigate how to develop responsible AI, aiming at fairness, accountability, transparency and ethical responsibility [9]. The importance of this topic is underscored by initiatives such as the new IEEE 7000 standard, to address ethical concerns during systems design, which advocates that intelligent systems should be ethically designed and responsible [10].

Responsible AI is inclusive. This means that the opportunities that AI provides should be accessible for all who might benefit, and should not disadvantage persons with specific (maybe atypical) characteristics. This notion corresponds with the World Wide Web Consortium's (W3C) goal of making the network-access benefits"available to all people, whatever their hardware, software, network infrastructure, native language, culture, geographical location, or physical or mental ability", and their W3C Accessibility Guidelines (WCAG) for making web content more accessible to users with disabilities. However, a recent review shows that there is still work to do as current web accessibility standards insufficiently guarantee equal access for conversational agents, like chatbots, and that there is a lack of a coherent set of design guidelines or recommendations [11,12]. Additionally, *responsible AI is transparent and explainable.* Stakeholders should have appropriate insights into the way decisions were made (so-called process-based explanations), why a system acts like it does (outcome-based explanations), and how they can influence the consequences (so-called actionable explanations) [13,14,15,16,17]. Consequently, AI's explanation processes should be inclusive too, i.e., be able to express and process explanations for all affected persons (direct stakeholders), including those with specific (including atypical) characteristics.

* Corresponding author
E-mail addresses: m.l.tielman@tudelft.nl (M.L. Tielman), mcsuarez@fi.upm.es (M.C. Suárez-Figueroa), arne.jonsson@liu.se (A. Jönsson), mark.neerincx@tno.nl (M.A. Neerincx), L.CavalcanteSiebert@tudelft.nl (L. Cavalcante Siebert).

https://doi.org/10.1016/j.techsoc.2024.102685
Received 15 December 2021; Received in revised form 9 July 2024; Accepted 14 August 2024
Available online 6 September 2024

The need for inclusive, transparent, and explainable AI for people with specific characteristics is closely related to Article 9 of the Convention on the Rights of Persons with Disabilities (CRPD). In this article "Accessibility" refers to "enabling persons with disabilities to live independently and participate fully in all aspects of life", which includes "to ensure to persons with disabilities access, on an equal basis with others, to information and communications, including information and communications technologies and systems". Another concrete example is the European Accessibility Act, which aims to achieve accessible products and services, by removing barriers for persons with disabilities and elderly people. We argue that these barriers also exist in AI systems which do not provide inclusive explanations. For instance, if an AI system provides a non-inclusive explanation of what data it will share, it will prevent some people from making informed decisions about what to share with this system.

The term 'explainable artificial intelligence' (XAI) is typically used without an exact definition in the existing literature. For this work, we adopt the definition that XAI is any type of AI system that in some way tries to convey information to stakeholders about what it does, how it does that and why. This includes both information about the inner workings of the system, as well as about the process of how the system was made (such as design goals, who it was made for, etc.). Most XAI is focused on direct users or developers, but it can also be targeted at other stakeholders when relevant, such as policy makers or users' families. What exactly constitutes an AI system is in itself difficult to define, as there is no clear consensus on what AI is. For this reason, we take the widest possible scope and focus on any type of system which in some way displays either human-like thinking or behavior; or rational thinking or behavior [18].

XAI has received a lot of attention in the last years [19]. The research started from a rather technical perspective. More recently, researchers advocated for a human perspective and proposed to include social science theories & methods in XAI research & development [20,21,22]. However, we identify that this research has not yet fully addressed the issue of inclusive XAI. This is in line with the limited attention that inclusive AI has received in general. Despite research on inclusive technologies in general being more prominent, research in AI has been more focused on using AI to assist people rather than making AI itself inclusive in terms of explainability. For example, Schouten and colleagues [23,24] designed and tested an AI-agent that teaches low-literates to participate in society (e.g., online banking, obtaining forms from the municipal counter) without a capability to explain its own teaching behaviors. Many different forms of exclusion exist, and XAI could exclude people for different reasons. For instance, offering only verbal explanations would exclude people with hearing problems, using cultural metaphors or concepts could exclude different cultural groups and long complex sentences could exclude people with cognitive disabilities. Different types of disability require different types of adjustments from technology, as well as from XAI.

Considering an explanation to be "the details or reasons that someone gives to make something clear or easy to understand", it is clear that XAI systems should provide understandable reasons for different aspects of their processes. This inherently means that XAI should take into account what their users can and cannot understand. Up to now, a major part of XAI research has focused on "opening the black box", often for the AI-developers and experts themselves. This means that explanations are often technically complex. Although an important first step towards XAI, many stakeholders impacted by AI systems are not experts in the field. Recently, social science contributions are being integrated into the research and development of explanations for less-than-expert users [25,26]. Several algorithmic approaches have been proposed to facilitate the creation of explanations tailored to diverse user groups [27], spanning from employing machine learning methods to generate a classifier that produces explanations alongside classifications [28], to enabling personalized explanations for a different user via user interaction [29]. This is a step in the right direction of making XAI more inclusive. However, very little attention has been spent on the needs of people with disabilities. We argue that given the goal of XAI to "make something clear to understand", special attention needs to be given to XAI for people with diminished cognitive abilities. In this regard, we refer both to (a) people with low literacy, that is, people with difficulties in processing verbal or textual information and (b) people with cognitive and intellectual disabilities.

People with cognitive disabilities have certain limitations in mental functioning. There are many different types of cognitive disabilities, which affect different cognitive domains. The World Health Organization (WHO) manual for assessment schedule evaluates cognition in the domains of concentration, remembering, problem solving, learning and communication [30]. The W3C Web accessibility initiative identifies cognitive challenges in memory, executive functions, reasoning, attention, language, knowledge and behavior [31]. These categories can also be seen in the cognitive ability domains [32], namely (a) language, communication, and auditory reception, (b) reasoning, idea production, and cognitive speed, (c) memory and learning, (d) visual perception, and, (e) knowledge and achievement [32], as well as in the categories of cognition aids which describes attention, memory, perception, decision and knowledge [33]. Although all slightly different, these categorizations all show the wide range of problems that can arise from cognitive disabilities. It is good to also mention low literacy. Although people with low literacy do not necessarily have a cognitive disability, many cognitive disabilities can lead to low literacy and given that many XAI methods focus on written explanations, low literacy is one of the most important user characteristics to take into account when designing inclusive XAI. For XAI to truly be responsible and inclusive, it is important that people with a range of different cognitive disabilities, too, are able to understand and appropriately trust the decision making of AI systems they interact with.

We have identified the need for responsible AI which is both inclusive and explainable, as well as a current research gap in the area of XAI for people with a cognitive disability. One of the reasons for this research gap is that the engineers working on XAI are often unfamiliar with the specific needs of people with different cognitive disabilities. And vice versa, researchers working with people with cognitive disabilities are typically unfamiliar with the inner workings of AI. We argue that to develop truly inclusive XAI, we need an interdisciplinary approach. Therefore, in this paper we present a roadmap for this research, identifying a vision, the state of the art in the relevant research fields, and a roadmap of how these fields can move forward to achieve this vision.

The methodology followed in this paper can be divided into four phases: (1) identifying the vision, where do we want to be; (2) identifying the research areas, what are the main research topics which are important in achieving this vision; (3) gaps identification, where are we now, and how does this fall short of our vision; and (4) roadmap definition, how do we start addressing the gaps. Our approach is inspired by previous research roadmaps and approaches [34,35]. We start by defining our vision in Section 2, in which we sketch what we wish to achieve. This description of the vision then helps us with the second step, in which we identify the important research areas by identifying the key questions that need to be answered to achieve our vision. This step we have added to our methodology because of the inherently interdisciplinary nature of the problem of inclusive XAI. It is important to look at what research fields are involved, which questions they need to answer and how they connect. This is done in Section 3. Once we know what main research areas and questions are important, we can describe the state of the art. Literature on the relevant fields was reviewed and summarized in Section 4, which helps with the identification of current gaps in making explainable AI inclusive to people with cognitive disabilities. In addition, links, connections and synergies among the aforementioned fields were discussed and identified. Finally, from this description of the field and the identified gaps, we achieve the roadmap to start addressing these gaps in Section 5.

2. Towards a general vision of inclusive explanation

In this section, we work towards describing our general vision of inclusive XAI for people with cognitive disabilities. The core of this vision is that XAI needs to generate explanations that are appropriate for the user and the situation they are in. It is important, therefore, to understand the nature of an explanation itself. To that end, this section first presents a conceptual explanation framework that provides a comprehensive understanding of the different aspects of explanations in the context of intelligent systems. The main goal of such a framework is to explain in a visual way the key concepts and their relations with respect to explanations; that is, to reach a general understanding of what explanations are on a conceptual level. The elements in the framework should be taken into account in the development of inclusive explainable intelligent systems. Following, we present our vision on inclusive XAI through two different scenarios describing how users could interact with an AI system and its explanations.

2.1. Conceptual explanation framework

Our conceptual explanation framework can be defined as an ontology and presented in a graphical way as an ontological model. We focus in this paper on the conceptual level of the framework; thus, we depict the most important concepts and relations in our framework as the conceptual model shown in Fig. 1 and 2.

To create this conceptual model we took as inspiration the NeOn Methodology for developing ontologies [36,37]. In this regard, we followed a modelling lifecycle structured in 4 main activities: (1) analysis of theory related to XAI explanations, (2) analysis of resources to be reused, (3) development of the model, and (4) application of the model to a pair of use cases.

Activity (1) included the study of research works in the area of XAI in order to better understand the key elements involved. This study allowed us to create a mental map of the crucial aspects when an explanation is needed in the context of AI systems.

Activity (2) has been performed taking into account the reuse-based approach, which is considered a best practice in Ontology Engineering as indicated in the NeOn Methodology. In this regard, we carried out the following tasks: (1) searching for candidate ontological models that could satisfy our needs; (2) assessing whether the candidate models are useful for our purpose; and (3) selecting the best candidate model for developing our conceptual explanation framework. Such tasks have been conducted considering all the aspects identified in Activity (1).

Activity (3) included the reuse of the selected model, that is, the conceptual overview of the Explanation Ontology [38,39], which encodes the system and user attributes of explanations that would allow them to be generated computationally. We extended this model with explanation levels, explanation modes, explanation types, explanation content, user types, and language level:

- Currently, we distinguish between two different explanation levels: (a) explanation about the process performed by the AI system and (b) explanation about the outputs provided by the system. These levels are described as individuals in the model.
- An explanation mode[1] can be of different types [40], thus we decided to consider the following ones: (a) Data Visualizations, which present data used in the AI system so the users can form their own understanding; (b) Cases or Scenarios, which present specific examples that support the AI system process and/or result; (c) Analytic (didactic) statements in natural language that describe the elements and context that support a choice; and (d) Rejections of alternative choices (or "common misconceptions" in pedagogy) that argue against less preferred answers based on analytics, cases, and data. The considered types are represented as instances in the model
- There are six different types of explaining AI decisions: (a) Rationale Explanation, which refers to non-technical reasons for taking a particular decision; (b) Responsibility Explanation that provides information about people responsible for the AI system; (c) Data Explanation, which informs about the data used for taking a particular decision; (d) Fairness Explanation that provides information about equality aspects in the taken decisions; (e) Safety and Performance Explanation, which refers to issues related to security and accuracy of the AI system; and (f) Impact Explanation, which informs about issues regarding the AI system's general impact.
- Regarding the elements that are used to describe the explanation content, we consider the following ones: (a) Natural Language Statements, (b) Images, (c) Graphics, (d) Icons, and (e) Pictograms. The set of content elements can be expanded, if needed. In the case of Natural Language Statements, different language levels can be used; those levels are Technical Language, Simplified Technical Language, Standard Language, Plain Language, and Easy-to-Read Language. These levels are represented as model instances.
- Users can be divided into these types (inspired from Ref. [41]): (a) Developers and (b) AI researchers who are involved in the development of the AI system; (c) Domain Experts, who are specialists in the AI system domain and participate also in the development, and (e) End Users and (f) End Users with Cognitive Disabilities, who use the systems. Additionally, we propose to add (g) Caregivers to this model. For some people with cognitive disabilities, caregivers such as family or medical professionals might be involved to help them make decisions. In this case, they might also become the user for whom the explanation is intended. We further acknowledge that within these user categories, it is important to recognize that there might still be important differences between users. For instance, users with different cognitive disabilities will require very different explanations depending on the nature of their disability.

Activity (4) can be seen as a kind of validation since we represented the knowledge related to explanations for a pair of use cases. We express such knowledge as natural language triples (<Subject Verb Object>).

Use Case 1: A medical diagnosis AI system runs a predictive model about the risk of a patient of getting a heart attack [42]

<User1 isConsumerOf Explanation1>
<User1 instanceOf EndUser>
<Explanation1 instanceOf RationalExplanation>
<Explanation1 hasContent Content1>
<Content1 instanceOf NaturalLanguageStatements>
<Content1.content=Given that the patient has high blood pressure the rist to have a heart attack is high>
<Content1 isWrittenWith PlainLanguage>
<User2 isConsumerOf Explanation2>
<User2 instanceOf DomainExpert>
<Explanation2 instanceOf DataExplanation>
<Explanation2 hasContent Content2>
<Content2 instanceOf NaturalLanguageStatements>
<Content2.content=Given that the patient's chest pain type is asymptomatic and the slope peak segment is upsloping, there ir a 89% probability that s(he) will get a heart attack.>
<Content2 isWrittenWith TechnicalLanguage>

Use Case 2: Credit approval [38]

<User1 isConsumerOf Explanation1>
<User1 instanceOf EndUser>
<Explanation1 instanceOf ImpactExplanation>
<Explanation1 hasContent Content1>
<Content1 instanceOf NaturalLanguageStatements>
<Content1.content=If co-signer with credit rating over 750, loan

[1] The explanation mode is different from the explanation modality, which mainly refers to the format. This format could be a specific document or a specific set of elements in the user interface of the AI application, to mention the most probable options.

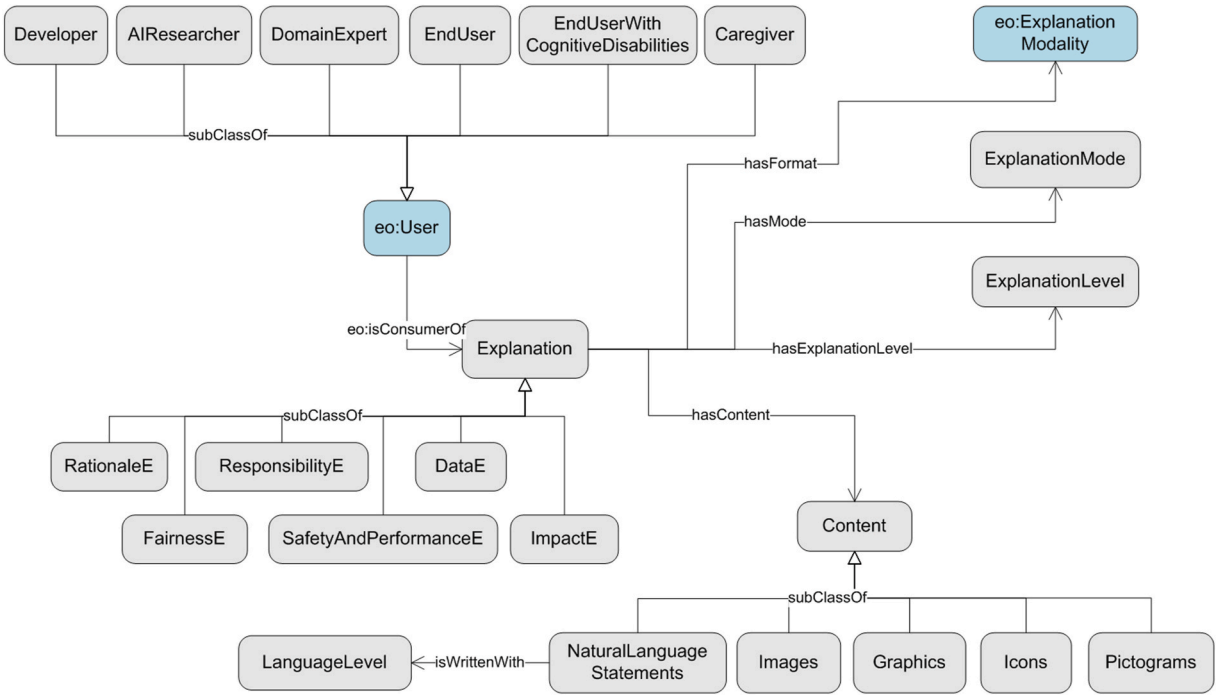

Fig. 1. Overview of the conceptual explanation framework.

would be more likely to be approved.>

<Content1 isWrittenWith PlainLanguage>

2.2. Scenario examples

Having established a conceptual understanding of explanations and what is involved, we can describe our vision for inclusive XAI. In the two scenarios below, we describe this vision using the personas of Mary and John. In both scenarios we illustrate the possible use of a future system, changing the focus from defining system operations (e.g., functional specification) to describing people's use and interaction with such a system following the scenario-based design principles [43]. This shift of perspective places the user and their needs at the center of the analysis.

These two scenarios highlight a number of important things. Firstly, they illustrate AI systems which should have a representation of and be able to reason about what is important for their users to know. Mary wants to know what data is used, but she doesn't really care how her route is determined exactly. It also acknowledges that sometimes, the user would wish for autonomy where it is difficult for the AI to fully provide all correct information. John, therefore, makes his privacy decisions with the help of others, and the AI can tailor the same information for different user groups. It is important to recognize that sometimes, simplification means a loss of information or nuance. In John's case, this means that other people help make the decision, whereas in Mary's case the AI just informs her about a loss of accuracy. Additionally, it shows the importance of tailoring explanations to the user, so of personalisation. Mary has problems with reading, so the app communicates verbally. John has problems with his memory, so the app communicates information in multiple formats and repeats it regularly. And the scenarios show that an AI should be able to explain different things. From what type of data is used, to how exactly decisions are made, or just how certain decisions are. Some of these are easier to achieve technically than others. In general, these stories show the need for AI systems which understand both what their users need to know, and which can use different types and modalities of explanations.

3. Key questions

The scenarios in the previous section show that the complexity of inclusive XAI lies in designing technical solutions which truly understand the needs of their users. This requires knowledge from both the field of technology for people with a cognitive impairment and the field of XAI, which are also both still fully in development. They present two sides of the coin of ever smarter systems; new opportunities (in how technology can help people), and new challenges (because this technology increasingly impacts us). Both these fields are already in themselves interdisciplinary, involving expertise from computer science, but also psychology, linguistics and philosophy. This means that to achieve our vision we need to understand how questions from different fields are

Scenario 1: Mary is a student with a reading disability. She can read a text written according to the Easy-to-Read (E2R) Methodology, but finds spoken interaction easier. She is interested in Prehistoric Artifacts, so visits a museum where she uses a dialogue system called ArtTour. ArtTour guides users while giving personalized information about the artifacts. To calculate the best route, ArtTour learns from past experiences and Mary's characteristics. It is important for Mary that she understands what data about her the system uses, but she does not really care how exactly the route is calculated. ArtTour explains to Mary verbally exactly which of her data it uses to generate the route. During the tour, ArtTour tells Mary what she is seeing verbally, slightly adjusting some difficult terminology to make it easier for her to understand. This information is often correct, but sometimes the easy version is factually slightly off. Mary cares about getting the right information. So whenever ArtTour suspects that this explanation might not be completely correct, it tells Mary verbally that this is the case, and gives her the option to hear the original, more complex information as well.

Scenario 2: John is 50 and has an intellectual disability that mainly affects his short-term memory. His doctor advised him to start eating more healthily for his health, so he installed the EatHealthy app to help him. The app uses personal data to advise him on what to eat using decision trees, and it has several data privacy settings. The more data the app has, the better its advice will be. It is important for John to agree with what data the EatHealty app collects, but it is difficult for him to make the trade-off between privacy and performance. Therefore, John wants to decide on the settings together with his family and the doctors. The EatHealthy app has generated three different written explanations of what data will improve its advice in what way, one for John, one for his family, and one for the doctors. The explanation for John also tells him when something is difficult to explain, so his family or his doctor might help to decide. During use, the app gives John written messages with simple advice, with a picture of an apple. For all advice, it tells John how certain it is of this advice in easy terms, using large icons with colors. The app also takes into account that John quickly can forget any advice or explanation it gives. So it regularly re-informs John that he decided on the privacy settings with his family and doctor.

relevant and how they relate to each other. In this section, we identify three key research questions related to these issues and demonstrate how they connect to our research topic.

Firstly, we need to ask: How explainable should a system be? We make the observation that different systems might require different levels of explainability depending on how they are used and what impact they might have. Sometimes it will be enough to offer a pragmatic explanation that the user's data is not shared. Meanwhile, in other cases, the system might use complex reasoning to determine which data is shared when, requiring the user to understand this reasoning itself. In our scenarios, for instance, we see that Mary wishes to know what data is used, but not how her route is determined exactly. This question about required levels of explainability is related to the discussion about human responsibility and control over AI systems. What do you as a user need to know to make informed decisions about using the system and what can be omitted and, hence, not necessarily addressed by XAI developers? The concept of meaningful human control has arisen from discussions on autonomous weapon systems, an application in which it is clear that decisions need to be accountable and thus explainable [44]. Explainability is envisioned to empower people to meaningfully control the behaviour of autonomous systems, instead of just being "in the loop". However, it seems clear that a system generating funny pictures might, for instance, require less explanation than one making life-or-death decisions. Moreover, for some cognitive disabilities, one might argue that it is more important that care-givers receive an explanation than the direct user. In our scenarios, we see this for John, where care-givers are involved because some of the decision making is too complex for him alone (note that shared, professional-client decision-making is an actual topic that XAI should address in the medical domain). These examples show that it is important to first know the goal of an explanation, to determine what you need a user to understand and in what level of detail or precision.

Secondly, we need to address the cognitive capabilities of the user, and how we can design explanations that fit those capabilities so that the maximum achievable understanding is reached. We need to ask: what level of understanding can the user reach, and what is the right type of explanation to help them reach this level? Even for people with no cognitive disabilities, an explanation suitable for, say, an engineer need not be the same as for a physician. Moreover, cognitive disabilities can be very diverse, and explanations need to suit the individual. You would not use the same type of explanation for someone with Alzheimer as for someone with low-literacy, or for someone with autism. The user's ultimate level of understanding can be seen as the product of the type of explanation on the one hand, and the basic knowledge and capability of the user on the other hand. For many users, a detailed mathematical understanding of algorithms might be beyond them even with explanations which are tailored to them. It is important to recognize the individual capabilities of the user, and to understand that their highest possible level of understanding can only be ever reached if the explanation matches their capabilities.

Finally, we need to ask: (how) can we implement an AI system which can generate the necessary explanations? We observe that different types of implementation have different potential for explainability. Some data-driven systems are more 'black-box' than others, and some systems rely more on knowledge-based techniques, which are inherently more easily explainable because of their more conceptual representations. Additionally, some systems build on techniques that allow them to explain certain aspects of decisions that are made. For instance, a data-driven model might still be able to point out which variables had the most influence on the decision making [45], while an agent-based system might be able to point out what goal was meant to be achieved, based on which beliefs the decision was made and which emotions can mediate the decision process [46]. These two examples would give different types of explanations, though. The way a system is implemented impacts what type and level of explanation is ultimately possible.

We propose that if and only if these three questions are answered, one can determine whether and how a specific AI system can be explainable in a given context to a specific person with a cognitive disability. To understand the relationship between these questions, we can use the metaphor of a pole vault. The required level of explainability sets the bar. The user is the vaulter, and the explanation is the pole. These need to match, the vaulter needs to be able to use the pole to reach the highest possible point, i.e. level of understanding. It is important to note that this might still not always be enough to vault the bar, even if it is the highest possible for this user. And the other way around, the highest possible level might not always be necessary if the bar is lower. Finally, the system is the manufacturer of the pole. To build a good one, it needs to understand what the user needs to jump the bar, and be able to create this. See also Fig. 2. Of course this is a simplified metaphor, true understanding can be about different topics or levels and cannot necessarily be expressed in a single value, but it does illustrate the relationships between these questions, and why all of them are necessary.

Given these three key questions, we know that three things need to be determined to develop inclusive XAI. Firstly, one needs to determine what level of understanding from the user would be necessary in the context of a specific system and application. Secondly, one needs to

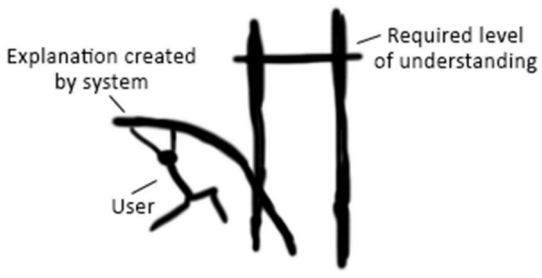

Fig. 2. A metaphor of the relationship between the key questions. The first question is about where to set the bar: how high the level of user understanding should be. The second question is about the optimal match between the user and the pole: what explanation does this user needs to achieve a level of understanding as high as possible. The third question is about how we manufacture that pole: how can a system create the necessary explanation.

understand the cognitive abilities of the user, and what type of explanation would be required for them to reach that necessary level of understanding. And if an explanation type exists that does this, the final question becomes how can we implement the AI system such that it can generate this type of explanation. We argue that these three questions are all crucial to provide good explainable AI for people with cognitive disabilities. These three main problems are tackled by different, though overlapping fields of research.

4. Where we are today

In the previous section, we described three main questions which we believe should guide research into inclusive XAI for people with cognitive disabilities. In order to present a roadmap towards the vision described in Section 2, we do not just need to know where we wish to go, but also where we are now. In this section, we present the current state of the art around research for inclusive XAI, led by the three main questions presented in Section 3.

4.1. How explainable should a system be?

To answer this question in a responsible and inclusive manner, we need to acknowledge that an AI system does not stand in itself [47,48]. First, it should be understood as a part of a socio-technical system, in which people will have different (and often conflicting) values, expectations, and even different cognitive abilities. Second, context matters: a system's expected level of explainability should correlate with the expected impacts of the use of the system in the context of operation. A system recommending a song requires less explainability than one recommending a medical treatment. Hence, defining how explainable a system should be requires the identification and evaluation of the expected (societal, legal, and ethical) impacts of the AI system in its socio-technical context.

We argue that a key element in defining how explainable a system should be depends on how much a human (e.g. user, developer, designer) can take responsibility for the identified system's impacts to themselves or others. For this, we turn to the concept of meaningful human control, which states that humans should ultimately remain in control of, and thus be responsible for the actions of systems with autonomous capabilities [44]. In Santoni de Sio & Van den Hoven (2018) two necessary conditions for meaningful human control, and hence for humans to be able to take responsibility, are proposed: 1) the system should track relevant human moral reasons relevant to the circumstances (tracking condition), and 2) the system should trace its action back to a proper moral understanding of one or more humans (tracing condition) [44].

The tracking condition requires a dual-relation: the system should track not only the environment but also the moral reasons of relevant humans. This requires identifying the relevant human agents and the relevant moral reasons at stake through a context-dependent analysis, considering all impacted stakeholders, in which the roles, needs, and responsibilities of the humans designing or interacting with the system are defined. The need to include stakeholders with different cognitive capacities makes this process even more challenging. Let us consider John's scenario. John wants to make privacy-related decisions when using EatHealthy together with his family and doctors. Hence, the system's explanations should not only consider John's personal views on privacy, but also take into account the opinions, reasons, and decisions he took together with his family and doctors. In other words, a system should provide explanations to John with respect to both his own reasons and to the reasons raised by family and doctors; but also provide explanations to the family and doctors with respect to John's reasons. In this manner, a system could still support John's autonomy while respecting decisions he wants to make together with others.

The tracing condition of meaningful human control requires a proper moral understanding of one or more humans involved in the system. Consequently, it is necessary to estimate people's capacity to achieve such moral understanding and thus to take responsibility. This understanding depends not only on effective communication and accessible information but also on one's cognitive abilities to achieve moral understanding. Fischer and Ravizza's theory of guidance control [49], on which the account of meaningful human control is based [44], argues that in order to be morally responsible for a given action, a person must have "taken responsibility" for the decisional mechanism. This condition aims to avoid situations where a person's decisional mechanism can be bypassed or not responsive enough, e.g., direct manipulation of the brain, phobias, drug addiction, but also cognitive disabilities. It has been suggested that individuals with mild intellectual disabilities can be eligible for taking responsibility by those they already find themselves emotionally engaged with, such as family, friends and caregivers [50]. In Mary's scenario we stated that she cares about getting the right information. For this reason, whenever ArtTour suspects an explanation might not be completely correct, it tells Mary verbally that this is the case, and gives her the option to hear the original museum explanation as well. Such personalized explanations, dependent on attributes such as one's cognitive abilities, context, and moral implications should emphasize the need for a person to understand the explanations provided and thus its moral implications.

Lastly, answering the question of how explainable a system should be must not be done in isolation by designers and developers only. To achieve such a complex balance between responsibility and (human) autonomy it is crucial that people with cognitive disabilities and caretakers are properly involved in participatory design processes from early on [51]. In this manner, people with cognitive disabilities (and their caretakers) can be properly informed of the cost-benefit trade-offs and limitations and align their expectations towards using such systems.

4.2. How can we best explain given cognitive disabilities?

The second main question to answer to achieve inclusive explainable AI is how to adapt explanations to fit the cognitive disability of the user. Just giving a 'general' explanation might miss the mark, leaving the user without the necessary understanding. However, this does not mean they will never be able to grasp what needs to be explained, it might just need a more tailored explanation.

Explanations can be given in different modalities and modes. How an explanation could be adapted to the cognitive level of a user, therefore, depends on the modality in which it is presented. Indeed, in some cases the choice of modality might even be important. Most current XAI research is focused on textual explanations, as people usually explain themselves in text. However, visual explanations are sometimes better to explain certain concepts, such as geographical information (e.g. 'where on the map do you need to go'). Depending on the cognitive disability of the user and what needs to be explained, either a visual or a textual format might suit better.

Although most focus in XAI has been on textual explanations, there is some work focusing on visual explanations. Most notably, visual explanations have been used to explain neural networks, in particular those working on visual data. For such systems, pictures can show what part of the original visual input was most important to make the decision, for instance Refs. [52,53,54]. A form of visual explanations can also be seen in research on shared mental models [55], which is often done in the context of human-AI teamwork. Schoonderwoerd et al. (2021) propose to develop reusable explanation design patterns, and provide a first set of graphical and textual patterns for health-care support systems [56]. Although typically not phrased in terms of explanations, often AI systems share visual maps of a common environment to show to the human what they know of the situation [57]. These maps explain something about the knowledge of the agent. In both these cases, these explanations are characterized by the fact that they are meant for expert users of the system. Knowing how to interpret the visuals typically requires at least some knowledge about how the system

works. There is a gap in using visuals for explanations to laymen users, let alone for people with cognitive disabilities.

In many cases, a textual explanation might be preferred above a visual one. Even if only because many AI systems do not have a graphical interface, but rather a conversational one. Much decision making is done from textual information, as can, for instance, be seen in the use cases. Thus, adapting explanations to a reader is important in order for them to make informed decisions. It is not feasible to have humans adapt all information for persons with cognitive disabilities, especially not explanations which are often generated in real-time. Therefore, we need services that automatically adapt explanations to make them easier to read and understand.

There is, however, not one way of adapting explanations that is suitable for all users; what works for John need not work for Mary and vice versa. The one size fits all approach to automatic text adaptation can perhaps be useful to some extent, but there will always be reader-specific issues that can not be neglected. For instance, it has been shown that persons with cognitive disabilities can find easy-to-read texts provided by public authorities to be difficult despite the fact that the texts were written with this target audience as one of their main addressees [58]. Furthermore, for individuals with cognitive disabilities, motivation is important for text comprehension, which means that adaptations that in different ways enhance the engagement of the reader could be equally important as purely linguistic modifications [59,60]. This is especially important when the information is used for informed decisions as for John and Mary in our scenarios (see Section 2.2).

The research on the intercept between readers and texts is scarce, and especially regarding the types of text simplifications that are helpful for readers with different types of profiles. As for studies concerning the extent of increased comprehension as an effect of text adaptations, the results are mixed. One study found that easy to read authority texts were especially difficult to read for persons with cognitive disabilities, indicating that the conducted simplifications were not useful [58]. Two studies have applied a set of (automatic) simplification operations and general easy to read guidelines on digital texts and showed that the modified texts increased reading comprehension for persons with cognitive disabilities [61,62]. However, some of the guidelines for the design of easy to read material have been questioned and further research is necessary.

This highlights the importance of involving the intended reader in the automatically simplified explanations, in order to make informed choices on how to adapt. Scenarios presenting interesting cases, like the Mary and John scenarios, may help to understand the tasks but give very little help in understanding their individual need for adaptation.

If an automatic explanation adaptation method or system targets a certain type of audience, the reader cannot be taken out of the equation. This end-user focus can of course take different forms. Perhaps the proposed method is derived from corpora with text written for a target audience, or the simplification operations are based on known psycholinguistic features. Perhaps the focus is instead manifested in the evaluation method.

Using data-driven methods for automatic adaptation, we must remember that our model can only be as good as our data. If we want to consider different target audiences, we should make sure that the data that we use for constructing our model mirrors the characteristics of explanations produced for that specific target group. To take a broad view, "simple text", will result in general models that do not take special needs into account. This does not imply that all such approaches are problematic, it has been shown that general simplification models can be very useful, but we should be aware of the limitations this might imply. A more general approach to how an explanation is simplified might need a thorough evaluation on participants from the target audiences, before we can say something about how they work for the targeted audience.

One source of information to understand how to adapt texts is corpora with texts written for audiences from different target audiences, especially for persons with cognitive disabilities, cf [63]. There are some, but not many such corpora. Most resources comprise texts written for the audience by professional writers, but there are also some sources that contain audience-specific data, such as common errors or gaze data.

We also need to assess the usefulness of decision making based on adapted explanations. Text complexity measures can be used to measure the complexity of the simplified explanation and compare that to the original text, cf [64]. Since automatic text adaptation systems for target readers ultimately aim to simplify explanations for a reader, the ideal way of evaluating such systems should be by including the specific audiences. However, such evaluations are time-consuming and it is not always easy to recruit participants from the different groups. Eye-tracking is a method that provides insights into the reading process of the individual, and possibly also the target audience, but the method is rather resource-demanding and to get enough participants to get a statistically sound basis for analysis might be a challenging task. Another drawback of using humans to assess the performance of simplification systems is that the results cannot be used for the comparison of other similar methods or systems.

4.3. How do we design and test a system capable of generating the right explanations?

The first part of this section gives a short overview of recent research and applications of inclusive design methods, which, in our view, will help to develop inclusive XAI in the near future. Subsequently, we will discuss the requirements for the XAI technology to generate the desired (actionable) explanations for all potential users, and argue that the XAI community would profit from a shared research focus on the development of inclusive XAI design patterns.

The Socio-Cognitive Engineering (SCE) method has been proposed and applied to research and develop human-centric AI, coherently addressing the social, cognitive and affective processes of human-AI collaboration and interaction [65]. This method explicitly addresses the individual differences in competencies and preferences to engage in these processes and, consequently, to develop technology that adapts to these differences. Techniques are provided to involve ageing people, people with low literacy skills and/or children in the development process of the information and communication technology, including qualitative research and analysis methods from anthropology to identify and interpret the user needs, values and varieties. Cremers et al. (2014) give an overview of these techniques with two example applications, following a grounded theory approach, identifying (1) usage-context and user-skill constraints for low-literates and (2) a value-model for child and parent for data-sharing (including value-tensions), which all should be mitigated by the technology [66]. Based on these models, adaptive agent technology was developed that addressed the specific varying user needs and usage contexts and assessed in the two case studies, children [67] and low-literates [23]. For the low-literates [24], derived a scaffolding model for an eCoach that learns low-literates practical skills to participate in society. For children it is noteworthy that they seem to prefer belief-based explanations somewhat more than adults, in relation to goal-based explanations [68], and that the specific context of the explanations can constrain children to benefit from it [69]. Taken together, these examples show that the SCE-method can help to identify "inclusion requirements" for explanations, and that such explanations should be grounded in a model of the user's competencies and preferences, and the context in which the associated social, cognitive and affective constraints appear.

By integrating XAI pattern engineering into the SCE-method, advancements in AI's explainability (i.e., the ability to deliver explanations [70], can be shared and re-used, e.g., Ref. [71]. Schoonderwoerd et al. (2021) present such an approach, aiming at an evolving XAI pattern library in the research & development community. The XAI design patterns explicate the design rationale with a multi-disciplinary grounding, establishing solutions for recurring design problems.

Furthermore, prototyping tools can be used to create and test the interaction components systematically [56](cf., [72,73]). The proposed pattern engineering approach has not yet been applied to develop inclusion solutions for recurring XAI design problems.

Looking at XAI technology, a rich set of XAI methods have been developed [74], from which LIME [75], SHAP [76] and MDNet [77] are the most common now. LIME and SHAP provide model-agnostic methods for additive feature attribution that try to approximate model's behavior or output (without accessing their internal variables). MDNet is more an "interpretable medical image network" for the diagnosis of cancer through the generation of textual diagnosis along with word-wise attention maps. Recent frameworks or ontologies, like the Unified Framework of [78] and the explanation ontology of [39], can be used to structure and relate the diverse XAI methods and underlying technology.

However, one crucial high-level requirement has received little attention till recently: The explanations should be actionable, i.e. the individual user (explainee) can process the explanations appropriately with his or her available momentary capacities (e.g., competence, self-efficacy), leading to the desired consequential behaviors. For this, the FACE (Feasible and Actionable Counterfactual Explanations) approach has been proposed that generates counterfactuals, "which are achievable and can be tailored to the problem at hand" [79]. Instead of the current XAI-practice, these counterfactuals address the nature of the target counterfactual and its real-life context by uncovering similar instances and feasible paths (e.g., advising a sport via example instances that represent the explainee's capacity to engage in this sport or develop the required skills). The FACE algorithm generates explanations that comprise actionable and feasible paths to meet a certain goal. It uses a f -distance quantification of the path-length and -density trade-off, and allows to impose additional confidence constraints. The FACE algorithm was applied to the MNIST data set [79], but not tested with end-users in a real-world use case. To establish and apply the proposed development of actionable explanations, we need data-sets that provide a sound representation of the (future) explainees, including persons with disabilities, and human-in-the-loop evaluations with these explainees.

Concerning this last requirement, human-centred evaluation methods and frameworks have recently been developed and proposed to be used by the XAI community [80,81]. For specific application domains, like health care, general participatory co-design methods have been instantiated and applied [82,56]. A next step would be to work out and apply dedicated techniques for inclusiveness and accessibility into these iterative design approaches.

In conclusion, there is a lack of real-world studies and developments in XAI, which address the XAI engineering process and the application of XAI technology for people with cognitive disabilities. Most of the research, so far, focuses on presenting "feature importance" and limited examples of explanations that reveal the causality for the Final User. Personalisation and context-dependency have been recognized as important requirements, but proven examples have been reported hardly. Contrastive and counterfactual explanations can support personalisation, by enabling individual choices of the "foil" or "counterfactual", but do not necessarily lead to feasible and actionable explanations. This section pointed out the research and development methods and technologies that can be applied to generate inclusive explanations. Sharing the concerning data-sets and design-solutions is needed to advance the work in this field in an efficient and effective way.

5. Research roadmap

The previous section described the state of the art in different fields of research which are important to design and develop more inclusive XAI for people with cognitive disabilities. Given the vision presented in Section 3, we can now start to identify how to approach this vision. This section will present a research roadmap towards this goal.

Firstly, we can see what the main research gaps are for each of the research fields from Section 4. When considering the question of how explainable a system should be, we see that the way forward is to consider the system not just as a technical entity, but within the broader socio-technical context. This means that the users, other stakeholders and context are crucial. This is because to answer this question, we need to understand the role of explanations in building moral awareness and enabling users to take responsibility. As a proper attribution of responsibility can only be determined if one understands the user's moral understanding of the decisions and consequences, we need to investigate the role of explanations in this moral understanding. Aside from understanding the users and the context in which the system is used, this also means studying the role of other stakeholders should the user's own moral understanding not be sufficient. Going forward, this gives us a number of new research questions. Firstly, how explanations of AI systems relate to the moral understanding of users. Secondly, how systems could start to understand and express the conditions for proper attribution of responsibility. And thirdly, how we can understand and model the role of other stakeholders and the duties associated with responsibility taking.

The second question addressed in Section 4 is how to best explain given a user's cognitive abilities. In this paper, we address this question mostly from the perspective of adapting explanations to the user's ability, with a focus on textual explanations. This is because visual explanations seem under-studied for simpler contexts and people with cognitive impairments, being used mostly in complex and visual domains. More work is necessary in this area, especially given the prevalence of low-literacy in this user group. Regarding text, we see that there is a need to focus on more than just simplicity in text. Rather, personalisation and diversification are the key. This can also mean that things like engagement are sometimes more important than simple language, depending on the user. Going forward, more focus should be given to this personalisation aspect, text should be personalized (addressing the situated social, cognitive, emotional and physical processes) rather than just simplified. This also means that the data these algorithms are based on should be diversified. And finally, that we need automatic feedback loops, i.e. systems that can measure whether this personalisation is successful (for instance through automatically measuring text complexity or engagement).

Finally, we addressed the question of how to design explainable systems. Broadly speaking, there are two parts to this question. Firstly, how to get the relevant information from an AI system (for instance about relevant features), and secondly about how to turn that information into an explanation. This section showed that while there is a lot of attention to XAI, it has been mostly on feature importance while the domain of causality remains more challenging. Going forward, generating information about causality will be crucial. Additionally, more research is necessary into how we can turn information into explanations. To do this, we need a socio-cognitive approach to XAI development, where the user is a part of the socio-cognitive system which needs to operate as a whole. The involvement of expert knowledge might be a way to help achieve this, as is establishing a common language about XAI. Viewing XAI as a social system is necessary for all XAI, but for inclusive XAI in particular given this user group's need for personalisation of explanations. Going forward, more research needs to be done into personalisation and memory.

Bringing all of these fields together is necessary to truly achieve inclusive XAI. One red thread through all of these topics is the need for personalisation. In deciding what the user needs to know, in catering to their cognitive abilities and in how to present what information in the explanation. Adaptive XAI is the key to inclusive XAI. We identify three main steps in achieving this:

1. Understanding what the XAI system needs to adapt to.
2. Performing the adaptation; change the explanation based on personal characteristics.
3. Evaluating the adaptation.

These three main steps describe the way forward. To understand what the XAI system needs to adapt to, we need to understand our user. The need for inclusive processes and data is key here. Users with diverse cognitive disabilities should be involved throughout the development process of XAI, and any data that is used should be representative of the user group. To perform the adaptation, we need the technical solutions to recognize the user and to know how to adapt explanations. This can be through machine learning methods, but could also be combined with expert knowledge and knowledge-based AI which allow for more conscious choices. Finally, evaluating inclusive XAI means that we need accountability and responsibility. We should have principles in place to check that the processes and data used in the design of such systems are indeed inclusive and do not cause harm to groups that are often already vulnerable. This requires logging and tracing of processes and algorithms. It also means evaluating systems in an inclusive way, actually involving the end-users also at this stage.

Finally, there is a need for a lingua franca, a common language across disciplines. When research fields come together, miscommunication is easy and the field of XAI is already very interdisciplinary. Common terminology, methodologies and understanding are necessary to bring this field forward. This is especially crucial as in inclusive XAI, the people developing the systems will typically not have first-hand experience with the user group, while the user group will not always know or understand the technical challenges involved. Shared language is the first step in bridging this gap.

6. Conclusion

In this paper, we address the topic of inclusive explainable artificial intelligence. We argue that with the rising interest in responsible AI, we need to focus on AI that is explainable in an inclusive way, especially for users with a cognitive disability. We identify that there is currently a lack of research on this particular topic, and address this by identifying a roadmap for research. This starts with a vision of XAI which is personalized to the needs, situations and capabilities of individual users with a cognitive disability. We have identified three main questions which need to be answered to achieve this; how explainable should a system be?; how do we craft an explanation which matches the capabilities of the user?; and how do we achieve such explanations in AI? From our overview of the current state of the art in these fields, we identify that the main ways forward lie in better inclusion of end-users in all steps of this process, techniques which can truly personalize AI explanations and accountability principles for inclusive XAI.

CRediT authorship contribution statement

Myrthe L. Tielman: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Mari Carmen Suárez-Figueroa:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Arne Jönsson:** Writing – original draft, Methodology. **Mark A. Neerincx:** Writing – review & editing, Writing – original draft, Methodology. **Luciano Cavalcante Siebert:** Writing – review & editing, Writing – original draft, Methodology.

Data availability

No data was used for the research described in the article.

Acknowledgments

The research described in this paper has been partially supported by the grant "Data 4.0 Project (TIN2016-78011-C4-4-R)" funded by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”. The participation of Myrthe L. Tielman and Luciano Cavalcante Siebert in this research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] A. Berman, K. de Fine Licht, V. Carlsson, Trustworthy ai in the public sector: an empirical analysis of a Swedish labor market decision-support system, Technol. Soc. 76 (2024) 102471, https://doi.org/10.1016/j.techsoc.2024.102471. https://www.sciencedirect.com/science/article/pii/S0160791X24000198.

[2] J.E. Rockoff, B.A. Jacob, T.J. Kane, D.O. Staiger, Can You Recognize an Effective Teacher when You Recruit One? Education Finance and Policy, vol. 6, 2011, pp. 43–74, https://doi.org/10.1162/EDFP_a_00022. https://direct.mit.edu/edfp/article/6/1/43-74/10196.

[3] C. Fontes, E. Hohma, C.C. Corrigan, C. Lütge, Ai-powered public surveillance systems: why we (might) need them and how we want them, Technol. Soc. 71 (2022) 102137, https://doi.org/10.1016/j.techsoc.2022.102137. https://www.sciencedirect.com/science/article/pii/S0160791X22002780.

[4] J. Kingdon, AI fights money laundering, IEEE Intell. Syst. 19 (2004) 87–89, https://doi.org/10.1109/MIS.2004.1. https://ieeexplore.ieee.org/document/1315546/.

[5] N. Rubens, D. Kaplan, T. Okamoto, E-Learning 3.0: anyone, anywhere, anytime, and AI, in: International Conference on Web-Based Learning, Springer, 2012, pp. 171–180.

[6] F. Burger, M.A. Neerincx, W.P. Brinkman, Technological state of the art of electronic mental health interventions for major depressive disorder: systematic literature review, J. Med. Internet Res. 22 (2020) e12599, https://doi.org/10.2196/12599. https://www.jmir.org/2020/1/e12599.

[7] M.S. Kannelønning, Navigating uncertainties of introducing artificial intelligence (ai) in healthcare: the role of a Norwegian network of professionals, Technol. Soc. 76 (2024) 102432, https://doi.org/10.1016/j.techsoc.2023.102432. https://www.sciencedirect.com/science/article/pii/S0160791X23002373.

[8] R. Williams, S. Anderson, K. Cresswell, M.S. Kannelønning, H. Mozaffar, X. Yang, Domesticating ai in medical diagnosis, Technol. Soc. 76 (2024) 102469, https://doi.org/10.1016/j.techsoc.2024.102469. https://www.sciencedirect.com/science/article/pii/S0160791X24000174.

[9] J. de Greeff, M.H. de Boer, F.H. Hillerström, F. Bomhof, W. Jorritsma, M. Neerincx, The FATE system: FAir, transparent and explainable decision making, in: Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), Palo Alto, CA, USA, 2021. http://ceur-ws.org/Vol-2846/paper35.pdf.

[10] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems", . Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Technical Report II. URL: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

[11] P. Silvennoinen, T. Rantanen, Digital agency of vulnerable people as experienced by rehabilitation professionals, Technol. Soc. 72 (2023) 102173, https://doi.org/10.1016/j.techsoc.2022.102173. https://www.sciencedirect.com/science/article/pii/S0160791X22003141.

[12] J. Stanley, R.t. Brink, A. Valiton, T. Bostic, B. Scollan, Chatbot accessibility guidance: a review and way forward, in: Proceedings of Sixth International Congress on Information and Communication Technology, Springer, 2022, pp. 919–942.

[13] J. Hangl, S. Krause, V.J. Behrens, Drivers, barriers and social considerations for ai adoption in scm, Technol. Soc. 74 (2023) 102299, https://doi.org/10.1016/j.techsoc.2023.102299. https://www.sciencedirect.com/science/article/pii/S0160791X23001045.

[14] Q.V. Liao, K.R. Varshney, Human-centered Explainable Ai (Xai): from Algorithms to User Experiences, 2021 arXiv preprint arXiv:2110.10790.

[15] M.A. Neerincx, J. van der Waa, F. Kaptein, J. van Diggelen, Using perceptual and cognitive explanations for enhanced human-agent team performance, in: Engineering Psychology and Cognitive Ergonomics: 15th International Conference, EPCE 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 15, Springer, 2018, pp. 204–214.

[16] C. Wilson, M. van der Velden, Sustainable ai: an integrated model to guide public sector decision-making, Technol. Soc. 68 (2022) 101926, https://doi.org/10.1016/j.techsoc.2022.101926. https://www.sciencedirect.com/science/article/pii/S0160791X22000677.

[17] C.T. Wolf, K.E. Ringland, Designing Accessible, Explainable Ai (Xai) Experiences. SIGACCESS Access. Comput, 2020, https://doi.org/10.1145/3386296.3386302, 10.1145/3386296.3386302.

[18] S.J. Russell, P. Norvig, Artificial Intelligence: a Modern Approach. Prentice Hall Series in Artificial Intelligence, third ed. ed., Prentice Hall, Upper Saddle River, 2010.

[19] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främling, Explainable agents and robots: results from a systematic literature review, in: AAMAS '19: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019, pp. 1078–1088.

[20] C. Baber, P. Kandola, I. Apperly, E. McCormick, Human-centred explanations for artificial intelligence systems, Ergonomics (2024) 1–15.

[21] R.R. Hoffman, T. Miller, W.J. Clancey, Psychology and ai at a crossroads: how might complex systems explain themselves? Am. J. Psychol. 135 (2022) 365–378.

[22] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[23] D.G. Schouten, F. Venneker, T. Bosse, M.A. Neerincx, A.H. Cremers, A digital coach that provides affective and social learning support to low-literate learners, IEEE

Trans. Learning Technol 11 (2018) 67–80, https://doi.org/10.1109/TLT.2017.2698471. https://ieeexplore.ieee.org/document/7915719/.
[24] D.G.M. Schouten, P. Massink, S.F. Donker, M.A. Neerincx, A.H.M. Cremers, Using scaffolding to formalize digital coach support for low-literate learners, User Model. User-Adapted Interact. 31 (2021) 183–223, https://doi.org/10.1007/s11257-020-09278-0. https://link.springer.com/10.1007/s11257-020-09278-0.
[25] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for Explainable AI: Challenges and Prospects, 2019 arXiv:1812.04608 [cs], http://arxiv.org/abs/1812.04608. arXiv: 1812.04608.
[26] T. Miller, But why?" Understanding explainable artificial intelligence, XRDS 25 (2019) 20–25, https://doi.org/10.1145/3313107. https://dl.acm.org/doi/10.1145/3313107.
[27] M. Chromik, A. Butz, Human-xai interaction: a review and design principles for explanation user interfaces, in: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18, Springer, 2021, pp. 619–640.
[28] M. Hind, D. Wei, M. Campbell, N.C. Codella, A. Dhurandhar, A. Mojsilović, K. Natesan Ramamurthy, K.R. Varshney, Ted: teaching ai to explain its decisions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 123–129.
[29] K. Sokol, P. Flach, One explanation does not fit all: the promise of interactive explanations for machine learning transparency, KI-Künstliche Intelligenz 34 (2020) 235–250.
[30] T.B. Üstün, N. Kostanjsek, S. Chatterji, J. Rehm, Measuring Health and Disability: Manual for WHO Disability Assessment Schedule WHODAS 2.0, World Health Organization, 2010.
[31] L. Seeman, M. Cooper, Cognitive accessibility user research, W3C First Public Working Draft 15 (2015).
[32] J.B. Carroll, others, Human Cognitive Abilities: A Survey of Factor-Analytic Studies, vol. 1, Cambridge University Press, 1993.
[33] A.C. McLaughlin, V.E. Byrne, A fundamental cognitive taxonomy for cognition aids, Hum. Factors 62 (2020) 865–873, https://doi.org/10.1177/0018720820920099. http://journals.sagepub.com/doi/10.1177/0018720820920099.
[34] G.I. Broman, K.H. Robèrt, A framework for strategic sustainable development, J. Clean. Prod. 140 (2017) 17–31, https://doi.org/10.1016/j.jclepro.2015.10.121. https://linkinghub.elsevier.com/retrieve/pii/S0959652615015930.
[35] J. Faludi, S. Hoffenson, S.Y. Kwok, M. Saidani, S.I. Hallstedt, C. Telenko, V. Martinez, A research roadmap for sustainable design methods and tools, Sustainability 12 (2020) 8174, https://doi.org/10.3390/su12198174. https://www.mdpi.com/2071-1050/12/19/8174.
[36] M.C. Suárez-Figueroa, A. Gómez-Pérez, The neon methodology framework: scenario-based methodology for ontology development, Appl. Ontol. 10 (2015) 107–145, https://doi.org/10.3233/AO-150145. https://content.iospress.com/articles/applied-ontology/ao145.
[37] M.C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, A. Gangemi, Ontology Engineering in a Networked World, Springer, 2012.
[38] S. Chari, O. Seneviratne, M. Ghalwash, S. Shirai, D.M. Gruen, P. Meyer, P. Chakraborty, D.L. McGuinness, Explanation ontology: a general-purpose, semantic representation for supporting user-centered explanations, Semantic Web Pre-press (2023) 1–31, https://doi.org/10.3233/SW-233282. https://content.iospress.com/articles/semantic-web/sw233282.
[39] S. Chari, O. Seneviratne, D.M. Gruen, M.A. Foreman, A.K. Das, D.L. McGuinness, Explanation ontology: a model of explanations for user-centered ai, in: International Semantic Web Conference, Springer, 2020, pp. 228–243.
[40] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.Z. Yang, XAI—explainable artificial intelligence, Sci. Robot. 4 (2019). Publisher: Science Robotics.
[41] M. Ribera Turro, A. Lapedriza, Can We Do Better Explanations? a Proposal of User-Centered Explainable Ai, 2019.
[42] R. Confalonieri, G.Guizzardi, On the Multiple Roles of Ontologies in Explainable AI, Neurosymbolic Artificial Intelligence, pre-print (2023), https://www.neurosymbolic-ai-journal.com/paper/multiple-roles-ontologies-explanations-neuro-symbolic-artificial-intelligence.
[43] M.B. Rosson, J.M. Carroll, Scenario-based design, in: Human-computer Interaction, CRC Press, 2009, pp. 161–180.
[44] F. Santoni de Sio, J. van den Hoven, Meaningful human control over autonomous systems: a philosophical account, Front. Robot. AI 5 (2018) 15, https://doi.org/10.3389/frobt.2018.00015. http://journal.frontiersin.org/article/10.3389/frobt.2018.00015/full.
[45] C. Agarwal, A. Nguyen, Explaining image classifiers by removing input features using generative models, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[46] F. Kaptein, J. Broekens, K. Hindriks, M. 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), IEEE, Lisbon, 2017, pp. 676–682, https://doi.org/10.1109/ROMAN.2017.8172376. http://ieeexplore.ieee.org/document/8172376/.
[47] V. Dignum, Responsible Artificial Intelligence: Designing AI for Human Values Publisher, Daffodil International University, 2017.
[48] M. Johnson, A. Vera, No AI is an island: the case for teaming intelligence, AI Mag. 40 (2019) 16–28.
[49] J.M. Fischer, M. Ravizza, Responsibility and Control: A Theory of Moral Responsibility, Cambridge university press, 2000.
[50] D. Shoemaker, Responsibility and disability, Metaphilosophy 40 (2009) 438–461, https://doi.org/10.1111/j.1467-9973.2009.01589.x, wiley.com/10.1111/j.1467-9973.2009.01589.x.
[51] S.K. Oswal, Participatory design: barriers and possibilities, Communication Design Quarterly Review 2 (2014) 14–19. Publisher: ACM New York, NY, USA.
[52] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable Ai: A Review of Machine Learning Interpretability Methods. Entropy 23, 18. Publisher, Multidisciplinary Digital Publishing Institute, 2021.
[53] Z. Qi, S. Khorram, L. Fuxin, Embedding deep networks into visual explanations, Artif. Intell. 292 (2021) 103435. Publisher: Elsevier.
[54] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, Journal of Imaging 6 (2020) 52. Publisher: Multidisciplinary Digital Publishing Institute.
[55] C.M. Jonker, M.B. Van Riemsdijk, B. Vermeulen, Shared mental models, in: International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, Springer, 2010, pp. 132–151.
[56] T.A. Schoonderwoerd, W. Jorritsma, M.A. Neerincx, K. van den Bosch, Human-Centered XAI: developing design patterns for explanations of clinical decision support systems, Int. J. Hum. Comput. Stud. (2021) 102684. Publisher: Elsevier.
[57] M. Demir, N.J. McNeese, N.J. Cooke, Understanding human-robot teams in light of all-human teams: aspects of team interaction and shared cognition, Int. J. Hum. Comput. Stud. 140 (2020) 102436. Publisher: Elsevier.
[58] L. Falk, S. Johansson, Hur fungerar lättlästa texter p\aa webben? Undersökning av lättlästa texter p\aa offentliga webbplatser. http://www.funka.com/contentassets/755a917870714fc8b9dd5a34de6a3237/rapport_lattlast.pdf, 2006.
[59] J.T. Guthrie, A. Wigfield, N.M. Humenick, K.C. Perencevich, A. Taboada, P. Barbosa, Influences of stimulating tasks on reading motivation and comprehension, J. Educ. Res. 99 (2006) 232–246, https://doi.org/10.3200/JOER.99.4.232-246. http://www.tandfonline.com/doi/abs/10.3200/JOER.99.4.232-246.
[60] A. Wigfield, J.T. Guthrie, Relations of children's motivation for reading to the amount and breadth or their reading, J. Educ. Psychol. 89 (1997) 420–432, https://doi.org/10.1037/0022-0663.89.3.420, apa.org/getdoi.cfm?doi=10.1037/0022-0663.89.3.420.
[61] I. Fajardo, V. Ávila, A. Ferrer, G. Tavares, M. Gómez, A. Hernández, Easy-to-read texts for students with intellectual disability: linguistic factors affecting comprehension, J. Appl. Res. Intellect. Disabil. 27 (2014) 212–225, https://doi.org/10.1111/jar.12065, wiley.com/10.1111/jar.12065.
[62] J. Karreman, T. van der Geest, E. Buursink, Accessible website content guidelines for users with intellectual disabilities, J Appl Res Int Dis 20 (2007) 510–518, https://doi.org/10.1111/j.1468-3148.2006.00353.x, wiley.com/10.1111/j.1468-3148.2006.00353.x.
[63] L. Feng, N. Elhadad, M. Huenerfauth, Cognitively motivated features for readability assessment, in: Proceedings of the 12th Conference of the European Chapter of the ACL, 2009. Edition: 229-237 event-place: Athens, Greece.
[64] M. Santini, A. Jönsson, E. Rennes, Visualizing facets of text complexity across registers, in: Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI), 2020, pp. 49–56.
[65] M.A. Neerincx, W. van Vught, O. Blanson Henkemans, E. Oleari, J. Broekens, R. Peters, F. Kaptein, Y. Demiris, B. Kiefer, D. Fumagalli, B. Bierman, Socio-cognitive engineering of a robotic partner for child's diabetes self-management, Front. Robot. AI 6 (2019) 118, https://doi.org/10.3389/frobt.2019.00118. https://www.frontiersin.org/article/10.3389/frobt.2019.00118/full.
[66] A.H.M. Cremers, Y.J.F.M. Jansen, M.A. Neerincx, D. Schouten, A. Kayal, Inclusive design and anthropological methods to create technological support for societal inclusion, in: D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, A. Kobsa, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, C. Stephanidis, M. Antona (Eds.), Universal Access in Human-Computer Interaction. Design and Development Methods for Universal Access, vol. 8513, Springer International Publishing, Cham, 2014, pp. 31–42. http://link.springer.com/10.1007/978-3-319-07437-5_4, 10.1007/978-3-319-07437-5_4. series Title: Lecture Notes in Computer Science.
[67] A. Kayal, W.P. Brinkman, M.A. Neerincx, M.B.V. Riemsdijk, A user-centred social commitment model for location sharing applications in the family life domain, IJAOSE 7 (2019) 1, https://doi.org/10.1504/IJAOSE.2019.106429. http://www.inderscience.com/link.php?id=106429.
[68] F. Kaptein, J. Broekens, K. Hindriks, M. Neerincx, The role of emotion in self-explanations by cognitive agents, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2017, pp. 88–93.
[69] F. Kaptein, J. Broekens, K. Hindriks, M. Neerincx, Evaluating cognitive and affective intelligent agent explanations in a long-term health-support application for children with type 1 diabetes, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, Cambridge, United Kingdom, 2019, pp. 1–7, https://doi.org/10.1109/ACII.2019.8925526. https://ieeexplore.ieee.org/document/8925526/.
[70] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018) 1–42. Publisher: ACM New York, NY, USA.
[71] K. Pollmann, D. Ziegler, A Pattern Approach to Comprehensible and Pleasant Human–Robot Interaction. Multimodal Technologies and Interaction 5, 49. Publisher, Multidisciplinary Digital Publishing Institute, 2021.
[72] E. Saad, J. Broekens, M.A. Neerincx, An iterative interaction-design method for multi-modal robot communication, in: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, 2020, pp. 690–697.

[73] A. Sauppé, B. Mutlu, Design patterns for exploring and prototyping human-robot interactions, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014, pp. 1439–1448.
[74] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, arXiv preprint arXiv:2006.00093 (2020).
[75] M.T. Ribeiro, S. Singh, C. 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 (KDD '16) (2016) 1135–1144, https://doi.org/10.1145/2939672.2939778.
[76] S. Lundberg, S. Lee, A Unified Approach to Interpreting Model Predictions, 2017. CoRR abs/1705.07874, http://arxiv.org/abs/1705.07874. arXiv:1705.07874.
[77] Z. Zhang, Y. Xie, F. Xing, M. McGough, L. Yang, Mdnet: A Semantically and Visually Interpretable Medical Image Diagnosis Network, 2017. CoRR abs/1707.02485, http://arxiv.org/abs/1707.02485. arXiv:1707.02485.
[78] S. Palacio, A. Lucieri, M. Munir, J. Hees, S. Ahmed, A. Dengel, XAI Handbook: towards a Unified Framework for Explainable AI, 2021 arXiv preprint arXiv: 2105.06677.
[79] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, FACE: feasible and actionable counterfactual explanations, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 344–350.
[80] I. Donoso-Guzmán, J. Ooge, D. Parra, K. Verbert, Towards a comprehensive human-centred evaluation framework for explainable ai, in: World Conference on Explainable Artificial Intelligence, Springer, 2023, pp. 183–204.
[81] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Measures for explainable ai: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance, Front. Comput. Sci. 5 (2023) 1096257.
[82] C. Panigutti, A. Beretta, D. Fadda, F. Giannotti, D. Pedreschi, A. Perotti, S. Rinzivillo, Co-design of human-centered, explainable ai for clinical decision support, ACM Transactions on Interactive Intelligent Systems 13 (2023) 1–35.