Explainable AI (XAI) in a societal citizen perspective – power, conflicts, and ambiguity

Dorthea Mathilde Kristin Vatn[1] and Patrick Mikalef[1]

[1] Norwegian University of Science and Technology (NTNU), Trondheim, Norway
dorthea.vatn@ntnu.no

Abstract. Explainable AI (XAI) as a research field has had a substantial growth in the last decade driven by a curiosity to investigate "the inside" of AI models appearing as black boxes through developing techniques and methods. While XAI mainly has been treated as a technical artefact, the human stakeholder perspective is clearly underscored in recent definitions of XAI, explicitly underscoring that XAI should be regarded as more than just techniques and methods. However, what a human stakeholder focus means in organizational and societal settings is currently unexplored. This paper therefore aims to explore how the concept of XAI can be applied in societal settings by first presenting a layered theoretical understanding of XAI, suggesting that societal explainability dimensions might be grasped through an analysis of the current discourse surrounding AI in public press. We use a sample of news articles published in the Norwegian public press early summer 2024 regarding Meta's approach to use personal data in training of AI models to perform a discourse analysis. The aim of the analysis is to provide insights into how the articles create a perception of reality relating to the future AI system Meta aims to develop. Our analysis reveals that the public is presented with oppositions constructing a reality surrounding the future AI system Meta aims to develop, with shifting power dynamics, conflicting interests, and ambiguity in responsibility as major themes present in the current discourse. We end by discussing the implications following from our analytical approach and findings.

Keywords: Explainable AI (XAI), sociotechnical theory, human-centered AI.

1 Introduction

The recent years' increasing access to data in combination with increased computational power have contributed to a growing development and use of Artificial Intelligence (AI) across domains (Mikalef & Gupta, 2021). While the early rule-based AI systems were easy to understand and explain, the recent development of more complex machine learning models have given an increased interest in terms like transparency, interpretability and explainability. As such, the research field of explainable AI (XAI) has had a substantial growth in the last decade (Arrieta et al., 2020). This growing interest has in many ways been driven by a technical aim to develop methods, tools and techniques providing insights into the decision-making process within AI models appearing as

black boxes. While a lot of the literature within the field of XAI has been very technically oriented about providing insight into the algorithmic black boxes, recent work underscores that XAI is a broader research field of interest beyond data science. For instance, Miller (2019) frames XAI as a field at the intersection of AI, Human-Computer Interaction (HCI), and social sciences. Recently it has also been underscored that XAI might be approached from *two* different angles (Weber et al., 2024). On one side you have those describing XAI as a narrow subfield only looking into AI models appearing as black boxes, while the other side has a broader view on how AI systems in general explain their decisions (Weber et al., 2024). In the latter approach, the human perspective is considered important, and "XAI targets all types of users thus entailing social and psychological aspects" (Weber et al., *p*. 302). Even though XAI as a field seems to have been primarily driven by a curiosity to investigate "the inside" of AI models appearing as black boxes through developing techniques and methods, recent definitions frame XAI a *process* encompassing both data and application, in addition to the focus on developing techniques and methods (Coussement et al., 2024).

The growth within the field has been fueled by an aim to drive a socially responsible, ethical, and legal development of AI. The EU AI Act frames both transparency and the principle of human oversight as important aspects, further underscoring the usefulness of XAI to facilitate a responsible development of AI (EU, 2024). The responsibility dimension relating to XAI is underscored by Arrieta et al. (2020) explicitly linking XAI as a field to the concept of "Responsible AI". To facilitate the responsible development, implementation, and use of AI, XAI might serve as an important field ensuring that people have *knowledge* of the AI systems they interact with, enabling a closure of a "responsibility gap" (Taylor, 2024). Therefore, XAI is closely linked to a focus on the human interacting with AI systems, reflected in definitions of XAI such as: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand" (Arrieta et al., 2020, p. 85). By framing the explainability of AI as dependent on a given audience, careful analysis of *who* is the audience of a given AI system is necessary to address its explainability. Also, being very reflective on what "explainability" actually means will be essential, to avoid the pitfall that researchers just use their own understanding of what a good explanation consists of (Miller, 2019). This points at the importance of explicitly grounding explainability concerns on existing literature within fields addressing explanations, such as philosophy and psychology (Miller, 2019).

While the importance of focusing on the human interacting with AI is established within the XAI literature (Adadi & Berrada, 2018; Arrieta et al., 2020; Miller, 2019), what a human stakeholder focus means in broader perspectives such as in organizational and societal settings is currently unexplored (Vatn & Mikalef, 2024a; Brasse et al., 2023; Faßbender et al., 2022). Building on an understanding that XAI about making AI understandable in an audience-sensitive manner, it is reason to ask what XAI would mean in a citizen perspective, such as in the context of AI systems built on data from social media platforms. Therefore, this paper aims to explore how the concept of XAI can be applied in societal settings and the overarching research question serving as basis for the paper is twofold:

How can XAI be explored from a societal perspective encompassing the human as a stakeholder in the role as a citizen (1), and how does this impact our understanding of XAI as a research field (2)?

To answer the research questions, we first present the theoretical lens on XAI that has been guiding our work. This theoretical lens is built on a layered sociotechnical perspective of XAI expanded with a layer addressing the societal impacts of explainability. We depart from an understanding of XAI as a layered sociotechnical concept, and that each of these layers impose different *approaches* to get access to relevant explainability dimensions for a given stakeholder group. To get access to explainability dimensions relating to a societal citizen perspective we use a sample of articles from Norwegian public press published early summer 2024 about Metas approach to the use of personal data in training of AI models. These articles are analyzed with a discourse analysis providing insights into explainability dimensions we propose are relevant in a societal citizen perspective on XAI. The aim of the analysis is to provide insights into how these articles create a perception of reality relating to the future AI system Meta aims to develop, and how this might inform explainability concerns that are relevant for a societal perspective on XAI. We end by discussing the implications of both our analytical approach and findings.

2 From a sociotechnical perspective on XAI to a societal perspective on XAI

A lot of the research done on the topic of XAI has been technically oriented residing in the field of data science and AI, with a focus on creating techniques producing explainable models while at the same time maintaining their performance levels (Adadi & Berrada, 2018). Recent definitions are expanding the understanding of XAI to be about more than just techniques. Coussement et al. (2024) define XAI as "the process that allows one to understand how an AI system decides, predicts, and performs its operations" (p. 1), pointing at methods (i.e. techniques) being a central dimension of XAI, but also *data* and *application*. Also, the user-focus is also made evident by several (Vatn & Mikalef, 2024b; Arrieta et al., 2020; Miller, 2019), and as such, XAI is closely linked to a focus on the human interacting with AI systems. This is reflected in definitions of XAI such as: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand" (Arrieta et al., 2020, p. 85), or "XAI is a research field that aims to make AI systems results more understandable to humans." (Adadi & Berrada, p. 52139).

The human perspective is also underscored by Vatn and Mikalef (2024b) in their layered sociotechnical conceptualization of explainability of AI built on the human, technology, organization (HTO) framework (Karltun et al., 2017). Explainability is in this conceptual framework presented from five different perspectives, ranging from "explainability as a technical field", "explainability as communication of inner workings of AI to humans", "explainability as a context-specific user requirement", "explainability as a mean to achieve organizational goals", as well as "explainability as a

process". By framing of XAI as a multidisciplinary, layered, and dynamic concept, Vatn and Mikalef (2024b) argue that explainability is not something fixed, but something that is dependent on level of analysis referring to whether you apply a technical, human, or organizational lens to explainability. While Vatn and Mikalef (2024b) end their understanding of explainability at the organizational impacts of XAI, it is also reason to explore the broader societal impacts of a concept such as XAI, looking into what explainability of AI would mean in a citizen perspective. In figure 1, we frame explainability in five dimensions, ranging from explainability of AI in a technical perspective, explainability in a human perspective, explainability of AI in an internal organizational perspective, explainability of AI in an external organizational perspective, and explainability of AI in a societal perspective by focusing on citizens. Next, we elaborate on what these different perspectives encompass by providing examples from current literature addressing XAI from these different perspectives.

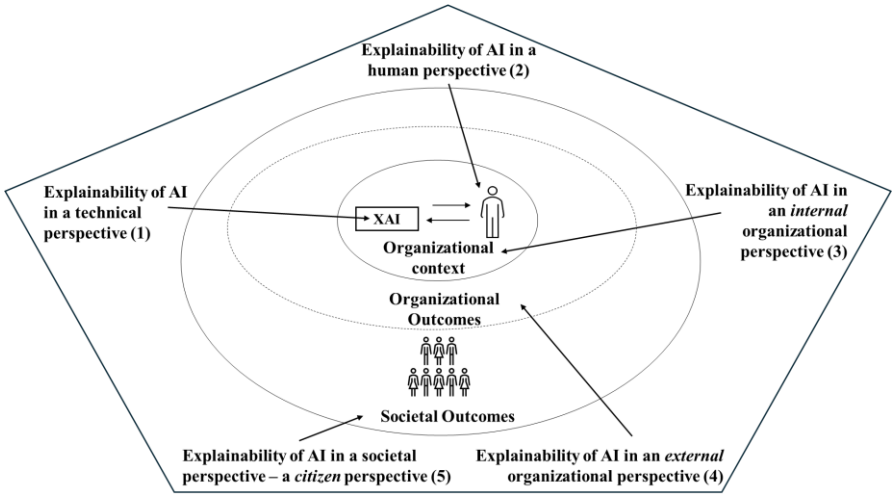

Fig.1. "Explainability" in five perspectives.

2.1 Explainability of AI in a technical perspective (1)

The technical dimension of explainability has up until now been the most prominent (Waardenburg & Huysman, 2022; Kim et al., 2024), and this perspective is underscored in definitions of XAI as a field concerned with creating a "suite of techniques that produce more explainable models whilst maintaining high performance levels" (Adadi & Berrada, p. 52138). While XAI as a research field has had a growth the last decade due to the development of more sophisticated models and methods (Arrieta et al., 2020), it is worth underscoring that the topic of explainability in the context of AI has been important for a long time (Weber et al., 2024; Sørmo et al., 2005; Gregor & Benbasat, 1999). Therefore, XAI from a technical perspective could also be regarded as more than techniques added to modern machine learning (ML) models appearing as black boxes. Arrieta et al. (2020) separates ML-models that are explainable due to being interpretable by design and ML-models that need external XAI techniques to give humans insight

into the prediction procedure. However, both of these categories are framed as XAI in ML by Arrieta et al. (2020), clearly underscoring the width in the understanding of XAI also from a technical perspective. In a technical perspective, XAI is also about extracting information from the AI-models useful for a human in the other end, but the "human" is often experts with deep knowledge of AI, such as AI experts, data scientists, and developers. In a technical perspective, explainability could therefore be regarded as techniques providing insights into the decision-making procedure in statistical terms (Adadi & Berrada, 2018), providing explanations in terms of numbers and equations. Therefore, it is first by explicitly focusing on the human generally as a recipient of the AI explanations that the scope of the audience is increased.

2.2 Explainability of AI in a human perspective (2)

The importance of the human when addressing the explainability of AI systems is underscored by several (Vatn & Mikalef, 2024, Miller, 2019, Arrieta et al., 2020, Adadi & Berrada, 2018). While Vatn & Mikalef (2024) underscores that a dimension of explainability is "communication of inner workings of AI to humans", Adadi and Berrada (2018) frames XAI as a "research field that aims to make AI systems results more understandable to humans" (p. 52139). Miller (2019) frames explainable AI as "an explanatory agent revealing underlying causes to its or another's agent's decision making (p. 2), and underscores simultaneously that a lot of research done on XAI seems to be based on researchers own intuitions of what makes good explanations on human terms. Therefore, Miller (2019) underscores the importance of building research into explainability of AI systems on fields explicitly treating explanations such as psychology, philosophy, and cognitive sciences. Looking into explainability in a human perspective could for instance encompass questions looking into how decisions made by AI systems should be conveyed so that they are aligned with knowledge into the human information processing process (Simkute et al., 2022; Zhang & Lim, 2022), pointing towards focusing on design and human-machine interfaces (HMI) as an entrance to explainability dimensions. While both explainability of AI in a technical and human perspective encompass important dimensions relating to XAI as a field, it is possible to add a further layer to explainability by adding considerations of context through an organizational perspective.

2.3 Explainability of AI in an internal and external organizational perspective (3 & 4)

While the organizational perspective on explainability of AI is not that much explored within the Information Systems field (Brasse et al., 2023), understandings of XAI focusing on stakeholders within organizational settings implicitly bring in an organizational perspective (Arrieta et al., 2020). Vatn & Mikalef (2024b) frame two dimensions of explainability relating to an organizational perspective by framing explainability both as "a context-specific user requirement" and "as a mean to achieve organizational goals". Arrieta et al. (2020) underscore that different stakeholder groups within the same organization might have very different explanation needs from the AI system,

ranging from the data scientists focusing on XAI as a mean to develop new functionalities, to managers and board members mostly concerned with explainability of AI as a mean to understand which AI solutions are in use within their organizations and the associated possibilities and risks. While explainability in an internal organizational perspective might be mostly concerned with explainability requirements connected to internal stakeholder groups using AI to perform different tasks, explainability in an external organizational perspective would focus on the outcomes of explainability, for instance in terms of business value (Vatn & Mikalef, 2024b). Having an external organizational perspective, point out the importance of focusing on a wide range of stakeholder groups operating in the landscape around a given organization, ranging from customers and clients to competitors and authorities. When aiming to grasp explainability dimensions relating to the organizational perspective, these could be about aspects both related to design of interfaces, but also knowledge into specifics relating to roles and responsibilities. In an external organizational perspective, relevant explainability dimensions would also relate to an AI system's compliance with laws and regulations.

2.4 Explainability of AI in a societal perspective – a citizen perspective (5)

While an organizational perspective on explainability opens for a focus on a wide range of stakeholder groups, it is first by addressing explainability in a societal perspective that explainability considerations concerning humans as part of the general public as citizens might be addressed. While an exhaustive mapping of all possible target groups of AI systems is challenging to provide, a usual generic framing of target groups could be in three categories focusing on developers, domain experts, and customers/clients (Faßbender, 2022). These target groups could be considered to play a role across all the different layers to varying degrees. However, citizens as a target group are a distinguishing aspect of the fifth layer presented in figure 1, and in the context of XAI, this perspective is not much explored (Faßbender, 2022). The lack of a citizen perspective might hinder a focus on explainability concerns for the public, and this is problematic as individuals meet AI as part of daily life, also outside organizational settings, for instance through social media platforms. As Faßbender (2022) points out that "(…) none of the traditional target groups are intrinsically concerned with *collective* interests and consequences", pointing towards the importance of addressing explainability concerns relating to citizens as a stakeholder group. What XAI actually is or would encompass from a citizen perspective is currently not explored, nor how to grasp explainability dimensions relevant in a societal citizen perspective.

Starting from a social constructivist standpoint, we suggest that a possible approach to examine explainability dimensions relating to the citizen perspective is to have a look at the current discourse surrounding AI in the current news press. By this we aim to explore explainability dimensions provided to the public by revealing how the language used in these articles creates a perception of reality relating to AI. While there are several definitions of the "discourse" term, a simplistic way of framing it is that it refers to the idea that language is structured in different patterns that our statements adjust to. As such, discourses both facilitate and constrain what can be said, by whom, in given

settings, and by looking into these patterns one can get a glimpse of how phenomena and worlds views are constructed (Willig, 2015). Next, we describe what we frame as the "Meta case" that unfolded early summer 2024 in Norwegian press, and how we used a Foucauldian discourse analysis of the news articles to explore how a societal perspective on XAI might be addressed encompassing the human as a stakeholder in the role as a citizen.

3 Case and methods

The topicality of addressing explainability from a citizen perspective is illustrated by Meta's plans on using personal data for training of their AI-models (Muhaisen, 2024). In May and June 2024 users of Meta's platforms Facebook and Instagram got a notification stating that they are updating their guidelines (Muhaisen, 2024), stating they will use the personal information shared by individual users to train their AI model. Instead of asking for consent, they provided information about individuals' right to oppose the use of personal data for this purpose. However, the process of opposing has by several been pointed out to be difficult to perform (Viken, 2024). The case resulted in several news articles and public debates, and in mid-June it became clear that Meta postponed their AI plans due to a request from the Irish Data Protection Authority (VG, 2024). This paper aims to explore how AI explainability dimensions relating to the citizen perspective might be grasped by using news articles to get a glimpse of the current discourse surrounding the future AI system Meta aims to build based on personal data shared by individual users on their platforms. Next, we describe how we practically used a sample of news articles to perform a discourse analysis.

3.1 Analytical approach

To explore explainability dimensions relating to the future AI system Meta aims to develop, we performed a Foucauldian discourse analysis (Willig, 2015; Jørgensen & Phillips, 1999). When performing a discourse analysis, the analytical focus is on the world views that the language use creates. The philosophical tradition of discourse analysis as a qualitative analysis method goes back to Ferdinand de Saussure and the study of signs and symbols and their use and interpretation. According to Saussure, a word is part of a network of other words it is different from, and it is by virtue of being different from other words that it acquires its meaning. Building on this tradition, the entry gate to a discourse analysis in its simplest sense is to carefully address how language creates an understanding of the world by revealing the oppositions present in the text. In our analysis of the news articles, we followed a process of firstly carefully reading through the articles underlining contradictions present within the news articles. These contradictions could be both implicit and explicit and are presented as oppositions with textual excerpts from the news articles in table 3 in the results section. Next, we grouped the oppositions presented across the different newspapers to form three themes summarizing our interpretation of the world views the news articles presented in relation to the AI system Meta aims to develop. Our analytical approach is built on a social

constructivist understanding of the world, meaning that the role of interpretation is a major aspect (Jørgensen & Phillips, 1999). Therefore, our analysis should not be regarded as an attempt to provide an objective answer to the research questions serving as basis for our work, but as a new contribution to the current discourse.

3.2 Sample of articles

The news articles that were included in the analysis were identified through following the news press as the case unfolded. As there was a vast amount of news articles published on the topic as the events unfolded, we ended up choosing six articles for analysis. The number of six articles was considered to enable a sufficient depth of the analysis, while at the same time provide insight into the unfolding events from the early articles elaborating on the statements given of Meta regarding their plans for this future AI system, to the late articles describing how the plans were postponed. In table 1 an overview of the included papers is provided, with information on both where and when it was published.

Table 1. News articles analysed

	Title of news article	Publisher	Publication date	Link
1	*Your most embarrassing party photos on Facebook can be used to train AI*	NRK	31st of May, 2024	https://www.nrk.no/norge/meta-skal-trene-ki-pa-dine-innlegg-og-bilder-1.16900169 (24.06.2024)
2	*Demands that the government calls Facebook [in for a meeting]- wants a ban*	TV2	2nd of June, 2024	https://www.tv2.no/nyheter/innenriks/krever-at-regjeringen-kaller-facebook-inn-pa-teppet-vil-ha-forbud/16733957/ (24.06.2024)
3	*Minister requests a meeting with Meta - reacts to AI plans*	VG	10th of June, 2024	https://www.vg.no/nyheter/i/Rzy6mJ/digitaliseringsministeren-ber-om-moete-med-meta-reagerer-paa-ki-planer (24.06.2024)
4	*The Minister of Digitalization gets to meet Facebook owner Meta*	Adressa	10th of June, 2024	https://www.adressa.no/nyheter/innenriks/i/25340r/digitaliseringsministeren-faar-moete-facebook-eier-meta (24.06.2024)
5	*Meta postpones controversial AI move*	VG	14th of June, 2024	https://www.vg.no/nyheter/i/KMpOyo/meta-utsetter-omstridt-ai-grep (24.06.2024)
6	*The Minister after meeting with Meta: - I emphasized that we have higher expectations of them*	Digi	17th of June, 2024	https://www.digi.no/artikler/statsraden-etter-mote-med-meta-jeg-understreket-at-vi-har-hoyere-forventninger-til-dem-br/548057 (24.06.2024)

4 A citizen perspective on XAI: power, conflicts, and ambiguity

The analysis of the articles gave rise to three themes; "shifting power dynamics", "conflicting interests", and "ambiguity in responsibility" that helped answer the research question on how XAI can be explored from a societal perspective. Table 2 provides an overview of the themes and how they are related to oppositions present in the articles.

Table 2. Themes and how they have emerged through oppositions in the news articles.

Theme	**Opposition**	**Excerpts from news articles**
Shifting power dynamics	Norway *versus* Big Tech	*"But Norway alone is unlikely to be able to influence the technology companies to implement it."* [2]
	The desire to oppose to Metas AI development *versus* being dependent on the platforms	*"Meta has made us all dependent on their infrastructure like Facebook and Instagram. It is not "just" to leave them"* [5] "*In practice, we are so dependent on a medium like Facebook that it is a decision that sits deep inside. It is certainly a problem that we are so dependent*" [2]
	Minister of Digitalization *versus* Representant of Meta	*"The Minister of Digitalization gets to meet Facebook owner Meta. Minister of Digitalization Karianne Tung (Ap) will meet a representative of Facebook owner Meta to talk about the plans to use people's photos for AI training*." [4]
Conflicting interests	Value of individual rights *versus* ensuring Meta's business interests by pointing at what "the others" do	*"Our approach is consistent with other leading technology companies, including Google and OpenAI"* [1] "*If there is a complaint and Meta has to withdraw its model, they risk two things (...). One is a fine (...). The second is to be required to delete their models or stop using them. - That is the worst thing that can happen to Meta because it will hit Meta's business model hard. I also believe that the solutions they create today can be the basis for future solutions."* [1]
	Societal need for technological development *versus* societal need for technology development built on trust	"*There is no doubt that data is a valuable resource to be used, and that we need AI that is trained in Norwegian and European languages and values, but that it is a prerequisite that data is collected in a way that builds trust."* [6] *"This is a step back for European innovation, competition in AI development and further delays bringing the benefits of AI to Europeans"* [5]
	Passing of time as beneficial for Meta *versus* sense of urgency on the governmental side	"*Meta is known to prolong the process as much as possible. When the decision finally comes, it may no longer be relevant*" [1] "*Time is running out, so the government must act now*." [2]
Ambiguity in responsibility	Individualistic data policy *versus* collective data policy	*"(...) she thinks they should ask for consent from the users before using our content for AI training.* *- If they don't do it, I was just as clear that I think they should make the reservation solution better and more easily accessible."* [6]
	Those who have technical competency *versus* those who don't	*"(...) especially worried about the users who are not tech savvy and believe many will not notice Meta's change at all. - For many, this change will go under the radar, and that is serious*" [1]
	Being comfortable with data sharing *versus* not being comfortable with data sharing	*"But I encourage people to refrain from the practice if they are not comfortable with it"* [3]

The news articles provide a world view where Metas AI plans are shifting power dynamics, are immersed in conflicting interests in a landscape where the responsibility of the individual versus the collective is ambiguous. Our approach of examining explainability dimensions through a discourse analysis of news articles provide explainability dimensions that might initially be regarded as quite far away from the explainability dimensions treated in the other perspectives (fig. 1). While it might be relatively easy to infer relevant explainability dimensions building on a traditional understanding of stakeholders surrounding an AI system encompassing the developer, the user, and the customer/client in a specific organizational setting, considering explainability concerns for citizens as a stakeholder group would require a different approach. Our analysis is built on an assumption that relevant explainability dimensions relating to the citizen perspective might be revealed by examining the current discourse surrounding a given AI system. Through our analysis of news articles concerning the Meta case we have revealed that the news articles provide a world view where Metas AI plans is shifting power dynamics, is immersed in conflicting interests in a landscape were the responsibility of the individual versus the collective system is ambiguous. By treating citizens as a stakeholder group in relation to this future AI system, we have revealed that the public is presented with a web of oppositions constructing a reality surrounding the future AI system Meta aims to develop (fig 2).

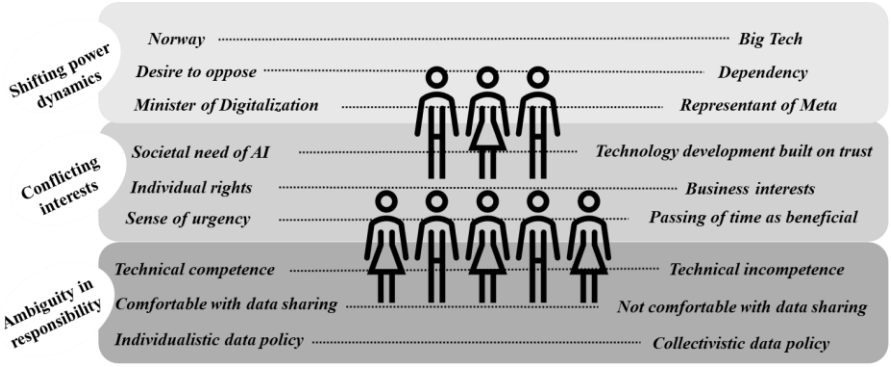

Fig. 2 Oppositions constructing a reality surrounding Meta's future AI system.

Next, we elaborate on how the emerged themes contribute to a social perspective on XAI encompassing the human as a stakeholder in the role as a citizen and discuss how this impacts our understanding of XAI as a research field.

4.1 Shifting power dynamics

A dimension visible in several of the articles relates to shifting power dynamics. As illustrated in the title "The Minister of Digitalization gets to meet Facebook owner Meta" [4], provide an impression that Meta is the powerful part, allowing the minister

to meet a representant "of the kingdom" of Meta. While carefully addressing the oppositions made in the articles, an evident opposition is made through the framing of Big Tech versus Norway [2], and "us versus Facebook" [3]. While outstanding politicians claim that it is obvious that Meta should ensure that users actively should consent to the use of their personal data to develop the AI, a powerlessness is made visible in the same politician's statement that "Norway alone is unlikely to be able to influence the technology companies to implement it" [2]. The impression of shifting power dynamics is also visible in the oppositions made between people on one hand feeling dependent on being on the platform [3], versus wanting to oppose the AI development plans.

4.2 Conflicting interests

A contradiction appearing in the news articles relates to Meta's consideration of rights of the individual users, versus the consideration of what "the other" big tech companies do [1]. By pointing at the practices of other tech firms like Google and OpenAI, a picture of reality of Meta as just continuing along a path established by others is created. This picture is also visible when looking at the oppositions created through the framing of business interests of Meta being in a direct conflict with the individual rights to personal information [1]. Also, by putting the need of Meta's AI services in opposition to our societal need of trust in data collection approach [5], reveal a conflict rising between the societal need to not fall "technologically behind", while at the same time not loose important societal values such as trust along the way. The aspect of time was important in several of the news articles presented [1,5], and the passing of time from the perspective of Meta versus the governmental side is framed as different. While Meta is framed to benefit from the passing of time due to both normalization and reduced relevance of legal concerns, the governmental side is framed to be in a state of urgency. By simply providing information that personal data will be used for training of their AI model such as Meta initially did, the copyright of the users' content is lost, and over time it contributes to a normalization of data use in ways that individual users don't have control over. As such ensuring consent from users in the use of personal data was put in opposition to the normalization happening over time of users' decreased rights to personal data. Also, the aspect of time was put in opposition to relevance, by pointing at Meta being known for interpreting and using laws and regulations to their own favor, pushing cases in the legal system as far as possible, and when a decision is made, it might not be relevant any longer [1]. This sense of urgency on the governmental side on the other hand was exemplified by a politician's statement: "Time is flying, the government needs to act now" [2]. For the reader this sense of urgency was also explicitly visualized in [1], providing the reader with a graphic illustrating countdown in days, hours, minutes, and seconds until the training of the AI of Meta would begin.

4.3 Responsibility ambiguity

While several public authorities are mentioned in the articles (e.g. Norwegian Data Protection Authority), an emphasis is made that it ultimately seems to be all about the individual user's choice [3], and an opposition is made relating to the feeling of comfort.

The framing of opposing to the AI plans of Meta as something the individual should do due to not feeling comfortable [3], underscores that the responsibility for opposing to the data practices serving as basis for the AI development, ultimately lies in the hands of the individual user. Linking this to the dimensions of technical competency brings in an important dimension relating to the awareness different users have of how the personal data is planned to be used in these AI models. This is illustrated in a statement made by a legal adviser in [1] stating that: "For a lot of people, this change will go under the radar, and that is serious". This points on technical competency and knowledge as decisive for how individuals actually use their choice to oppose the use of personal data for this AI system. Looking into current research literature on XAI, XAI techniques have been suggested to facilitate the closure of a "responsibility gap", ensuring that individuals might be responsible decision-makers also when interacting with future AI systems (Taylor, 2024). Despite XAI techniques' ability to provide human decision-makers with certain understandings of a given AI-system, Taylor (2024) asks if explainable AI truly is responsible AI. He concludes that XAI techniques have several limitations to serve as tools for ensuring responsibility. This points at the usefulness of looking more broadly on XAI than purely techniques to understand how it relates to responsibility, and in a citizen perspective an important question for the future is how citizens are should be educated, explained and trained in topics related to AI.

5 Conclusion and outlook

In this paper we have explored how XAI can be investigated from a societal perspective encompassing the human as a stakeholder in the role as a citizen, and challenged how this impacts our understanding of explainable AI as a research field. Addressing explainability of AI in a broader societal citizen perspective requires moving beyond the understanding of XAI as purely technical techniques providing insight into black boxes, or as simple design considerations. Through our approach we have therefore challenged our current understanding of XAI. In this paper, we have suggested that explainability of AI in a societal citizen perspective is a fifth layer in a conceptual understanding of explainable AI. We have suggested that relevant explainability dimensions in a societal citizen perspective might be identified by analyzing the current discourse in news articles. We have discussed how our approach would affect our understanding of explainable AI as a field, being fully aware that we are moving the understanding of XAI quite far from an understanding of it being about specific techniques. However, supporting ourselves on the notion that XAI should be regarded as a multidisciplinary field, we have begun an effort exploring how explainable AI can be conceptualized and analyzed from a societal citizen perspective.

Our analysis has revealed that the public is presented with a web of oppositions constructing a reality surrounding the future AI system Meta aims to develop, pointing at shifting power dynamics, conflicting interests, and ambiguity in responsibility as major themes present in the current discourse surrounding a future AI system that most citizens most likely would interact with. Taking the implications of our findings further, it is reason to ask what follows from a societal discourse painting a picture of AI system

development as changing power dynamics, revealing conflicting interests, as well as ambiguity in responsibility. Where are the boundaries between the individual's ability and right to decide for themselves how their data should be used for AI development by Big Tech, and the need for a collective governance of these processes? The field of explainable AI has grown last decade (Arrieta et al., 2020), and been closely connected with responsible AI. It is however reasonable to claim that the societal perspective on explainable AI currently is not much explored, despite its evident importance (Faßbender, 2022). Future efforts aiming to further explore the perspective of citizens as a stakeholder group in relation to XAI is therefore very welcome.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intel ligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, *33*(1), 26.

Coussement, K., Abedin, M. Z., Kraus, M., Maldonado, S., & Topuz, K. (2024). Explainable AI for enhanced decision-making. *Decision Support Systems*, 114276.

EU. (2024). *EU AI Act: first regulation on artificial intelligence.* https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (29.07.2024)

Faßbender, J. (2022). *Why explainable AI needs such a thing as Society*. https://www.hiig.de/en/explainable-ai/ (29.07.2024)

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, 497-530.

Jørgensen, M. W., & Phillips, L. (1999). *Diskursanalyse som teori og metode*. Frederiksberg: Roskilde Universitetsforl. Samfundslitteratur.

Karltun, A., Karltun, J., Berglund, M., & Eklund, J. (2017). HTO–A complementary ergonomics approach. *Applied ergonomics*, *59*, 182-190.

Kim, M., Kim, S., Kim, J., Song, T.-J., & Kim, Y. (2024). Do stakeholder needs differ? – De signing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces. *International Journal of Human-Computer Studies*, *181*, 103160.

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & management*, *58*(3), 103434.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artifi cial intelligence*, *267*, 1-38.

Muhaisen, S. (2024, 31st of May). De flaueste festbildene dine på Facebook kan bli brukt til å trene KI. *NRK*. https://www.nrk.no/norge/meta-skal-trene-ki-pa-dine-innlegg-og-bilder-1.16900169 (24.06.2024)

Simkute, A., Surana, A., Luger, E., Evans, M., & Jones, R. (2022). XAI for learning: Narrowing down the digital divide between "new" and "old" experts. In *Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference* (pp. 1-6).

Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, *24*, 109-143.

Taylor, I. (2024). Is explainable AI responsible AI?. *AI & SOCIETY*, 1-10.

Vatn, D. M. K., & Mikalef, P. (2024a). Explainable AI and business value – an organizational perspective. *Proceedings of the 32nd European Conference on Information Systems (ECIS)*. Cyprus.

Vatn, D. M. K., & Mikalef, P. (2024b). Theorizing XAI - a layered concept being multidisciplinary at its core. *Proceedings of the 30th Americas Conference on Information Systems (AMCIS)*, Salt Lake City.

VG (2024, 14th of June). Meta utsetter omstridt AI-grep. VG. https://www.vg.no/nyheter/i/KMpOyo/meta-utsetter-omstridt-ai-grep (24.06.2024)

Viken, O. (2024, 12th of June). Marit ville reservere seg mot Metas KI-trening – fikk svar på tysk. *NRK*. https://www.nrk.no/nordland/marit-ville-reservere-seg-mot-metas-ki-trening-_-fikk-svar-pa-tysk-1.16916596 (24.06.2024).

Waardenburg, L., & Huysman, M. (2022). From coexistence to co-creation: Blurring boundaries in the age of AI. *Information and Organization*, *32*(4), 100432.

Weber, R. O., Johs, A. J., Goel, P., & Silva, J. M. (2024). XAI is in trouble. *AI Magazine*, *45*(3), 300-316.

Willig, C. (2015). Discourse Analysis. In J. A. Smith (Ed.), *Qualitative Psychology; A practical guide to research methods*. (3rd ed., p. 143-167).

Zhang, W., & Lim, B. Y. (2022, April). Towards relatable explainable AI with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-24)