Two Means to an End Goal: Connecting Explainability and Contestability in the Regulation of Public Sector AI

TIMOTHÉE SCHMUDE, University of Vienna, Faculty of Computer Science, Research Network Data Science, Doctoral School Computer Science, Austria
MIREIA YURRITA, Delft University of Technology, Faculty of Industrial Design Engineering, Utrecht University, Faculty of Science, The Netherlands
KARS ALFRINK, Knowledge and Intelligence Design, Delft University of Technology, The Netherlands
THOMAS LE GOFF, i3, Télécom Paris, Institut Polytechnique de Paris, France
SEBASTIAN TSCHIATSCHEK, University of Vienna, Faculty of Computer Science, Research Network Data Science, Austria
TIPHAINE VIARD, i3, Télécom Paris, Institut Polytechnique de Paris, France

Explainability and its emerging counterpart contestability are key normative and design principles for trustworthy AI, enabling users and subjects to understand and challenge AI decisions. Yet realizing these principles is difficult, as they take on different meanings across technical, legal, and organizational dimensions of AI regulation. To address this conceptual polysemy, we report findings from an interview study with 14 experts examining the intersection and implementation of explainability and contestability, and their interpretations in different research communities. We outline differentiations between descriptive and normative explainability, judicial and non-judicial channels of contestation, and individual and collective contestation action. We also identify key points of friction in realizing both principles, including alignment between top-down and bottom-up regulation, assignment of responsibility, and the need for interdisciplinary collaboration. Finally, we offer three AI policy recommendations to operationalize explainability and contestability through a Regulation-by-Design perspective. Our contributions inform policy research and regulation of these core principles, and support more effective and equitable design, development, and deployment of trustworthy public AI systems.

Additional Key Words and Phrases: explainability, contestability, trustworthy AI, algorithmic accountability, transparency, AI governance, regulation-by-design, AI policy, interdisciplinary collaboration, public sector AI, qualitative methods

ACM Reference Format:
Timothée Schmude, Mireia Yurrita, Kars Alfrink, Thomas Le Goff, Sebastian Tschiatschek, and Tiphaine Viard. 2026. Two Means to an End Goal: Connecting Explainability and Contestability in the Regulation of Public Sector AI. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 27 pages. https://doi.org/10.1145/3805689.3812309

Authors' Contact Information: Timothée Schmude, timothee.schmude@univie.ac.at, University of Vienna, Faculty of Computer Science, Research Network Data Science, Doctoral School Computer Science, Vienna, Austria; Mireia Yurrita, m.yurritasemperena@uu.nl, Delft University of Technology, Faculty of Industrial Design Engineering, Utrecht University, Faculty of Science, Delft, Utrecht, The Netherlands; Kars Alfrink, c.p.alfrink@tudelft.nl, Knowledge and Intelligence Design, Delft University of Technology, Delft, The Netherlands; Thomas Le Goff, thomas.legoff@telecom-paris.fr, i3, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France; Sebastian Tschiatschek, sebastian.tschiatschek@univie.ac.at, University of Vienna, Faculty of Computer Science, Research Network Data Science, Vienna, Austria; Tiphaine Viard, tiphaine.viard@telecom-paris.fr, i3, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812309

1 Introduction

Explainability and contestability are central principles in the development and deployment of trustworthy public AI systems [118]. Explainability serves the purpose of increasing stakeholders' understanding of an AI system and to enable informed decision-making [64], while contestability allows stakeholders to challenge and appeal algorithmic decisions [73]. A range of methods ('mechanisms' [4]) has been suggested in explainable AI (XAI) and HCI through which these principles are realized, such as explaining feature importance [94], providing counterfactuals [110], or requesting human intervention [6]. But two main aspects remain underexplored: i) how explanation and contestation mechanisms intersect, and ii) how to proceed when implementing these mechanisms according to regulatory provisions.

While these principles are defined and discussed in both explainable AI (XAI) [52, 70, 116] and legal [6, 57, 77] research, an approach connecting these perspectives has not yet been adopted. The need for an empirical, unified approach is evident when considering the treatment of both principles in current AI regulation, such as the General Data Protection Regulation (GDPR) and Digital Services Act (DSA). While these texts can be understood to provide for explanations in AI systems to ensure contestability[1], there is an absence of guidelines to translate their legal provisions into tangible system requirements into practice [77]. With this work, we address this gap by contributing an empirically-grounded, qualitative analysis of how the principles of explainability and contestability can be connected and put to practice in public sector AI systems.

We formulate two challenges that impede the implementation of explainability and contestability in practice. First, both explainability and contestability are polysemic – they take multiple meanings depending on context – and require differentiation, as XAI, design, and legal research all employ the same terms but do not necessarily refer to the same concepts. Furthermore, the concrete realization of both principles depends on the involved actors [53], domain [111], and use-case setting [15]. This multiplicity of meaning and realization excludes a one-size-fits-all approach [43, 87] and instead requires guidelines that can be applied in a variety of contexts [77]. And second, both the theoretical background and regulatory oversight of AI in high-stakes scenarios is still in development. Consequently, there are few best practices and guidance that can aid in the implementation of contestability [5, 57, 73] and interdisciplinary approaches to the creation of legislation have only begun to be mapped out [83]. Closing this gap between regulation and implementation requires policy-making that is evidence-informed [75], i.e., that is supported by research that bridges disciplines and provides *empirical grounding*. To this end, our work is guided by the following research questions:

(1) *Conception*: What is the linkage between explainability and contestability in AI systems, and how does this manifest across conceptual, technical, and regulatory dimensions?
(2) *Implementation*: How can policy-makers support the implementation of explainability and contestability in AI systems?

Our findings highlight distinctions between descriptive and normative explainability, judicial and non-judicial contestation channels, and individual and collective contestation action. According to participants, its capacity for citizen empowerment is a key feature of connecting explainability and contestability; participants furthermore stressed this intersection as instrumental to both explainability and contestability's effectiveness and highlighted that both principles are not effective if the underlying policy governing a system's deployment follows values that are incommensurable with those of trustworthy AI.

The contributions of this work are three-fold: We first elicit empirical insights from experts in AI policy to contribute a detailed conceptual differentiation of explainability and contestability for public AI systems with respect to their goals, mechanisms, and challenges. We further provide a detailed analysis of the points of friction in the principles' implementation, emphasizing the spirit of the law, regulator responsibility, and a lack of

[1]The right to explanation is debated [21, 103, 110], but texts such as the GDPR, DSA, and EU AI Act provide for explainability in algorithmic decisions [77].

collaboration between communities. Lastly, we formulate three recommendations for AI policy, focusing on the potential of public deliberation and internal AI governance measures. We envision this work to provoke more evidence-informed [20, 75] discussion about the policy and implementation of explainable and contestable public AI systems.

1.1 Structure of the paper

The remainder of the paper is structured as follows: In Section 2, we give context for why the two principles of explainability and contestability matter for AI policy and introduce the concept of Regulation by Design [90]. In Section 4 and 5 we report on our study findings with respect to the intersection of both principles and their implementation in practice. In Section 6, we formulate these insights into recommendations to realize explainability and contestability through the lens of Regulation by Design. We give the study's limitations in Section 7 and conclude in Section 8.

2 Why connecting explainability and contestability matters for AI policy

Explainability and contestability are two means to an end goal [77]. This end goal is the development and deployment of trustworthy AI systems that preserve decision subjects' fundamental rights [41], support human agency and oversight [30], and adhere to the principles of procedural justice such as transparency and outcome control [65]. However, implementing the high-level principles can prove challenging in practice [39, 74], which is why fostering a perspective that contains both policy and design considerations is essential. Because of the interdependence of legal frameworks, design practices, and institutional governance, connecting explainability and contestability in both policy and design is an essential step in realizing Regulation by Design. In the following, we situate our work in relation to public sector AI (Section 2.1) and introduce working definitions of explainability and contestability (Section 2.2). Finally, we introduce a definition of regulation and the dimensions in which it is realized, situating our work within the Regulation by Design framework (Section 2.3).

2.1 Trustworthy AI systems in the public sector

AI systems deployed in public institutions can significantly impact individuals' fundamental rights, safety, or health [55]. Research shows that these systems are frequently dysfunctional [61, 93], discriminatory [22], and harmful through aggravating power imbalance [35, 84] and restricting autonomy [91]. For these reasons, both researchers [8, 12, 54, 97] and policy makers [36, 51, 86, 87] have advocated that high-risk AI systems should align with value frameworks such as *trustworthy* or *responsible* AI [25, 42, 106], which emphasize human agency, oversight, transparency, accountability, and fairness [41]. Explainability and contestability support these frameworks by enabling people to understand [64] and challenge [49] AI decisions. Although both principles are integrated into various design frameworks [5, 53, 64, 70], their implementation as part of the EU AI regulation remains challenging [40, 77, 83]. We used public sector AI as an "entry point" as we needed contexts where a body takes automated decisions affecting individuals' rights. Public sector AI provides this and enables us to ground the discussion in the interviews. However, the conclusions drawn from these case studies and analysis may not be specific to public sector AI and could thus be extended to all kind of high risk AI systems.

2.2 Explainability and contestability in policy, regulation, and design

Our work in this paper seeks to understand explainability and contestability from a multifaceted viewpoint that combines policy, design, and technology perspectives. Moreira et al. [80] use a systematic literature review to explore the relations between the two principles, understanding contestability as one of multiple aspects that factor into XAI and proposing an assessment scale to rate an ADM system's contestability. In contrast, this work focuses on an empirical exploration of the two principles grounded in the experiences and knowledge of

regulation and design experts, aiming to understand "the multiple gatherings and entanglements through which worlds of design, practice and policy are brought into messy but binding alignment" [56] – a form of policy work that is not usually documented in academic publications [113] and thus inaccessible through literature reviews.

2.2.1 From trustworthiness, through accountability and transparency, to explainability and contestability. As one of the most widely referenced frameworks across AI policy, law, and design communities, the EU HLEG[2] Guidelines offer a shared vocabulary for interdisciplinary work on explainability and contestability. The Guidelines define "trustworthy AI" as (1) lawful, (2) ethical, and (3) robust (technically and socially) [51]. AI systems are treated as sociotechnical systems [112], with trustworthiness ensured across the lifecycle [32]. Of the seven requirements, we focus on transparency and accountability in the connection of explainability and contestability.

Transparency requires information about data and models [51], decomposed into traceability (records of data/processes), communication (disclosing machine vs. human interaction), and *explainability* (accounts of technical and human processes behind decisions). When they have significant impact, decisions must be explainable "to the extent possible" to directly and indirectly affected persons, covering both local (individual decisions) and global (organizational processes, design choices, deployment rationale) explanations. Where outputs cannot be explained due to technical limits (e.g., black-box models [97]), other measures (e.g., traceability, auditability) are required. *Accountability* supports fairness by assigning responsibility across the lifecycle, operationalized via auditability, minimizing and reporting negative impacts, and redress, i.e., accessible mechanisms [88] for remedy, which we treat as individual, local contestability.

Trustworthiness depends in part on transparency, transparency requires explainability to decision subjects, and explainability supplies information that enables contestation. We offer working definitions of explainability and contestability on this basis, noting that they are contested: related terms (e.g., transparency, interpretability) vary across disciplines [48, 101]. Our definitions reflect a broadly shared understanding in the HLEG framework and adjacent policy and design literature and serve as starting points for our empirical inquiry, as the concepts' polysemy is itself central to this work.

2.2.2 Explainability. Explainability means providing *intelligible* information about an AI system's logic, core parameters, and specific purposes [64, 78]. 'Good' explanations are expected to adapt to stakeholders' expertise and objectives [64, 78], be it evaluating fairness, understanding the deployment context, or contesting a decision [5, 31, 100]. We follow this stakeholder-oriented understanding of explainability, which foregrounds the communicative function of explanations over mere information disclosure [78]. This definition differs from technical notions of transparency that concern whether a model's internal mechanisms are humanly graspable [67] – under which, for instance, exposing all parameters of a model could count as transparency but would not constitute an explanation in our sense – and from arguments for inherently interpretable models over post-hoc methods [97]. Although explainability is provided for in many policy texts [7, 37, 38], these texts leave implementation details unspecified, such as whether to use global or local explanations and whether to provide them before or after decisions [77].

2.2.3 Contestability. Contestability describes how various stakeholders, from human operators to decision subjects and third parties, can challenge algorithmic decisions [4, 6, 71] through data input control, decision revision requests, and various audit methodologies [65, 70, 104]. While the term 'contestability' is well-established in design and HCI research [4, 52, 71], legal scholarship more commonly refers to related concepts such as 'redress' or 'the right to contest' [58]. In both design and policy frameworks, the principle of contestability is described to be enabled through explainability by previous work [6, 77]. Contestation rights are contained in the EU Charter of Fundamental Rights [14], in GDPR Article 22, and the Council of Europe's Convention 108+[3] [6, 26].

[2]High-level expert group on artificial intelligence

[3]The Council of Europe's Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data.

2.2.4 Limitations of explainability and contestability. Current limitations to explainability and contestability implementation in AI systems stem from both technical constraints, policy challenges, and legal ambiguity.

From a *technical* perspective, creating workable explainability is not only challenging but even impossible for AI systems whose internal workings are fundamentally opaque to human observers, i.e., black-box models [62, 92, 97]. Post-hoc methods can extract local decision boundaries [94], feature contributions [69], and build surrogate models [79], but they exhibit fidelity gaps across subgroups that can cause fairness issues [9] and can yield conflicting attributions for the same prediction [13]. We thus follow Rudin [97] and [16] that post-hoc interpretations only approximate the model, are often unreliable, and can be misleading. Additionally, without benchmarks for explainable AI, user evaluations suffer from knowledge gaps across differing user needs [92]. Contestability also faces the challenge that much of today's AI is "static," offering no interactive interrogation of, or updates to, decisions [66].

From a *policy* perspective, explainability demands a workable reconciliation between explainability's technical and commercial limitations and an array of public governance standards [62]. Commercial limitations include the reluctance of tech firms to disclose trade secrets or other confidential information through explainability as well as the increasing costs of adhering to AI regulation [62]. Contestability suffers from governance processes that are closed to effective public contestation [62] and furthermore requires that policies openly specify the regulations and determinations of adequacy for specific applications. These include proper attribution of accountability and effective methods of compensation when decisions are challenged [3].

From a legal perspective, the GDPR right to explanation as a remedy for algorithmic opacity and unfairness has been debated [33, 109], with legal scholars noting ambiguity in the right's application and insufficiency in the technical focus of explanations. Recent CJEU cases, including *SCHUFA Holding* [27, 28] and *Dun & Bradstreet Austria* [29], clarified that GDPR Art. 22 applies whenever an algorithm plays a "decisive role" and requires controllers to explain the procedures and data used. While these rulings address scope and content, they do not resolve technical limitations or whether explainability effectively mitigates algorithmic unfairness.

2.3 Regulation by Design: from regulatory intent to institutional governance

This section defines the regulation process and introduces Regulation by Design (RBD), which we use to derive policy recommendations from our empirical insights. Regulation is the "sustained and focused attempt to alter the behavior of others according to defined standards or purposes" to produce a broadly defined outcome [10]. While classical regulation relies on public norms and command-and-control [19], modern regulation trends toward RBD [81, 114]: embedding regulatory objectives directly into technological architectures so that design choices govern systems [90]. Here, regulation is a rule-making activity enacted through social practices, and design is a situated social practice that alters environments. Unlike law or social norms, which act through incentives and sanctions, design can disable non-compliance, making it a uniquely powerful form of regulation. Figure 1 depicts the interdependence of legal frameworks, design practices, and institutional governance.

To illustrate, the rule "a valid ticket is needed to board a train" can be enforced by law (fines), norms (peer disapproval), or design (a turnstile blocking entry without a ticket) [89]. Design can prevent non-compliance. But design must itself be governed: law may require turnstiles to allow emergency evacuations, overriding the rule for safety. Governance thus meta-regulates how design regulates [89]. Design practices also shape norms and law: standardization can turn widespread designs into norms, and sometimes binding legal rules, e.g., "harmonized standards" under the AI Act [46]. Regulation must accommodate evolving technologies, values, and ethics while avoiding obsolescence, often via co-regulation in which authorities, developers, and users translate abstract legal norms into practice. This creates a feedback loop where regulatory intent and operational realities co-evolve [89].

For explainability and contestability, realizing RBD requires establishing congruent cross-disciplinary terminology and deepening mutual understanding of the practical problems faced by policy-makers and designers, a central concern of this paper.

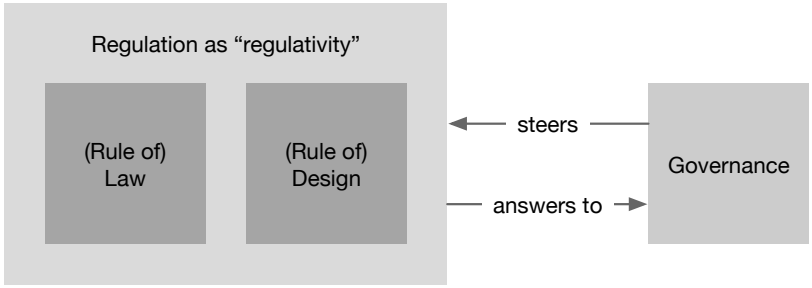

Fig. 1. Conceptual model of the relationship between design, 'classical' regulation (i.e., law-making), and governance. Both design and law-making are forms of regulation (i.e., 'regulativity'). Governance steers how design and law-making regulate, and design and law-making are both accountable to (and reflexively aware of) governance.

3 Methods used

We conducted task-based interviews with 14 experts in the regulation and design of AI systems. Participants' background included 6 legal scholars, 5 computer science scholars, and 3 scholars of social sciences; 5 participants were policy practitioners by occupation. Participants were recruited from European universities and public institutions.[4] Participants were screened for knowledge of legislation and implementation of AI policies and for their experience in their field of occupation (a list of screening criteria is included in the appendix). All participants were working on AI-related regulation and were actively involved with the topics of the study in their current role, such as through occupations in the EU institutions, standardization bodies, or legislative bodies. As this study aims to elicit details of how explainability and contestability are discussed and understood in policy-making circles, a form of policy work that is not documented in academic publications [113], we focused on sourcing insight from policy experts directly, rather than conducting a literature review. For participant recruitment, we relied on the authors' networks, snowball sampling, and direct invitations. This study uses a qualitative approach that aims for the exploration and construction of meaning to create "analytic rather than statistical generalization" [76]. The sample size thus followed qualitative research guidelines, focusing on code and meaning saturation [50].

3.1 Study procedure

Figure 2 gives an overview of the overall study process. All interviews were conducted online using the collaboration platform *Miro* and took around 90 minutes. The full interview guide is included in the appendix.

The interviews consisted of three key tasks: a card sorting activity, a use case discussion, and a discussion of interdisciplinary perspectives prompted through a citation network graph. Participants first sorted 40 cards containing explanation and contestation concepts and mechanisms into self-defined categories[5] to express their understanding of both principles. Participants then read a description of the French welfare fraud detection algorithm [95, 96] and reflected on the use of explanation and contestation from the perspectives of a fictional welfare beneficiary and the social security agency. Details on the use case and citation graph discussion are provided below. Lastly, participants explored an interactive network representation of academic literature on explainability and contestability.

[4]To allow participants the freedom of anonymous expression in the interviews, demographic as well as occupational details were omitted.
[5]The set of cards was created from literature on the explainability and contestability of AI systems, details are provided in the appendix.

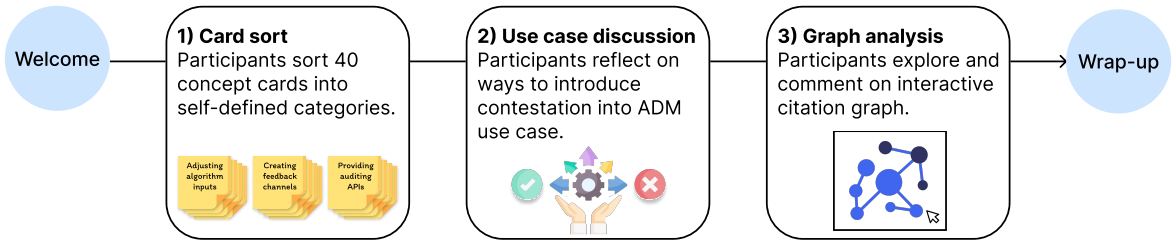

Fig. 2. **Study procedure overview.** Participants completed three sequential tasks: 1) sorting 40 concept cards into self-defined categories related to AI systems, 2) analyzing explainability and contestability in a specific algorithmic use case from personal and institutional perspectives, and 3) exploring an interactive citation graph of research publications on explainability and contestability.

These tasks enabled a three-part investigation of the RQs. Tasks 1 and 3 addressed RQ1's conceptual and regulatory linkages between explainability and contestability: the card sort elicited participants' mental models of categories and commonalities/differences [98], and the citation-graph interaction elicited how both principles are represented across disciplines (e.g., HCI, law) and where missing links indicate conceptual disconnects. Task 2 examined the implementation dimension (RQ1) and elicited policy advice (RQ2) via a concrete application scenario presenting two stakeholder inquiries on implementing explainability and contestability. The scenario included information and appeal issues from both decision subjects and a public institution, allowing participants to draw on their experiences with real-world cases. Themes emerging across tasks were then investigated in more detail through interview questions.

Interviews were recorded, transcribed, and analyzed using inductive *thematic analysis* [17], drawing on a method positioned between reflexive and reliability-oriented approaches [18]. An initial set of codes from the first five interviews were compared and consolidated between two authors to create a shared coding framework. The complete set of codes was then developed into themes capturing participants' conceptual understanding of explainability and contestability, their experience with the implementation of AI regulations, and their perspectives on the relation between institutions and decision subjects. The appendix includes the codebook as well as depictions of the concept cards and citation graph.

3.2 Use case

The study's use case presented a public AI system deployed by the Caisse d'allocations familiale (CAF), a part of the French social security services [96]. The system uses logistic regression to predict welfare fraud likelihood among beneficiaries and was trained on data from household investigations and corresponding overpayments. Around 13 million households (nearly half of France's population) are risk-scored by this system, of whom around 100,000 are flagged for detailed investigation per year [95]. In October 2024, civil society organizations criticized the system for alleged discrimination against marginalized groups and ineffectiveness [1], sparking public debate on high-risk AI systems and their regulation under the EU AI Act. The use case thus exemplifies broader challenges of integrating public AI systems into society and invites reflection on how explainability and contestability can support their resolution.

3.3 Citation graph

To elicit participants' perspectives on the meaning of explainability and contestability in different disciplines, they were asked to explore a web-based literature graph and to think-aloud during the interaction. To create this

network representation, we extracted 648 academic articles related to contestability as well as their references from the search engine Web of Science using the following query:

```
(ai OR "artificial intelligence" OR "algorithm* decis*") AND
("contest*" OR ("right to" AND "explain*"))
```

This query captured academic papers that mentioned both AI *or* algorithmic decisions *and* contestability *or* the 'right to explain' [103], creating links with explainability in legal literature while excluding unrelated work in XAI. This search was conducted in January 2025 and returned 648 papers, spanning from 1991 to 2025. However, 151 papers (23.3%) were published in 2024 alone and 465 (71.8%) were published in 2020 or after. The network G=(V,E) was built from these results by representing articles as a set of nodes V and inserting an edge between two articles if they cited at least k=4 of the same references. We set k manually after iterating through values between 1 and 10, selecting the value that increased visual readability. To exhibit clusters in the graph, we used the Louvain algorithm [11]. The graph is available under contest.graphuzu.fr and depicted in the appendix.

4 Understanding the intersection of explainability and contestability of AI systems

This section addresses our first guiding question (RQ1) by mapping the intersection of explainability and contestability of AI systems with respect to their conceptual, design, and policy dimensions. We use evidence from the task-based interviews and quotes from the participants (referred to as "P") to differentiate both principles and list their goals, mechanisms, and challenges. We first map how participants defined explainability and contestability (Section 4.1) to draw out differences in their understanding. We then show that a consensus emerged among participants on the fact that the two concepts follow the same goals while using different mechanisms (Section 4.2). We delineate how the intersection between the two notions is complicated by disciplinary gaps and divergences in Sections 4.3 and 5.

4.1 Differentiating explainability and contestability

Explainability is connected to two main notions: 1) understanding the technical workings of an AI system, such that "*the human user [...] is not treating anymore the machine as an oracle*" (P2); and 2) understanding the norms and reasons governing the AI's decisions, deployment, and institutional embedding, such as "*know[ing] who are the people in charge or who I can contact to give more information*" (P1). We define the first kind as *descriptive explainability* and the second kind as *normative explainability* or *justifiability* [63].[6]

Contestability means allowing stakeholders to challenge AI decisions, such as enabling regulators to "*critically process the information provided to them and push back against it*" (P6) and affected persons to "*understand the situation and to file complaints*" (P6). We define two key distinctions in how contestation is realized: The first distinction is between *collective action* and *individual action*. While individual contestation allows decision subjects to challenge AI decisions that affect them, collective contestation means joining with other stakeholders to mount a more general and broader contestation effort, i.e.,"*translating personal issues into general matters and public fights*" (P12). The second distinction is between *judicial channels* and *non-judicial channels* of contestation. Judicial channels use formal means provided by the judicial system to contest decisions in court, colloquially described as "lawyering up" (P9), while non-judicial channels support issue resolution through design solutions or direct human intervention, such as mediation systems or ombudspersons. Differentiating both the the type of contestation action and the channels used is essential to pinpoint the meaning of contestability: "*For legal scholars, [...] there is this idea of centering judicial proceedings and centering the courts, even though most of what we could call the 'life of the law' is not usually in the courts*" (P9).

[6] "Justification" has different meanings across disciplines. In XAI, it refers to descriptive rationales for model predictions, while in legal scholarship it concerns demonstrating decisions' lawfulness against norms. We distinguish explanations as descriptive accounts from justifications as normative ones grounded in external standards [49, 63].

4.2 Explainability and contestability follow the same goals but use different mechanisms

4.2.1 Goals. Participants repeatedly stated that explainability and contestability work towards the same goals and purposes. These goals can be summarized as supporting the rule of law and empowering citizens by alleviating opacity and information asymmetry: "*We cannot contest [if] we don't know what is used, to what purpose, and how it works*" (P3). Explanations provide descriptive and normative information that support assessing both the acceptability of individual decisions and the justifiability of a system's deployment. This dual assessment guides contestation at individual or collective levels. Acceptability is crucial when subjects face unfavorable outcomes and must decide whether to contest; as P9 noted, it helps ensure citizens use contestation channels *"not just to throw a cog in the wheel and delay procedures that you know that are going to be inconvenient to you, but are otherwise acceptable"* (P9). When individual contestation is not possible, explanations can prompt collective contestation – e.g., when underlying policy is unfair or dysfunctional. Explanations thus help stakeholders gauge their stance toward an AI system locally and globally and select appropriate contestation channels, determining whether contestation is between a decision subject and an officer or between a collective and the institution.

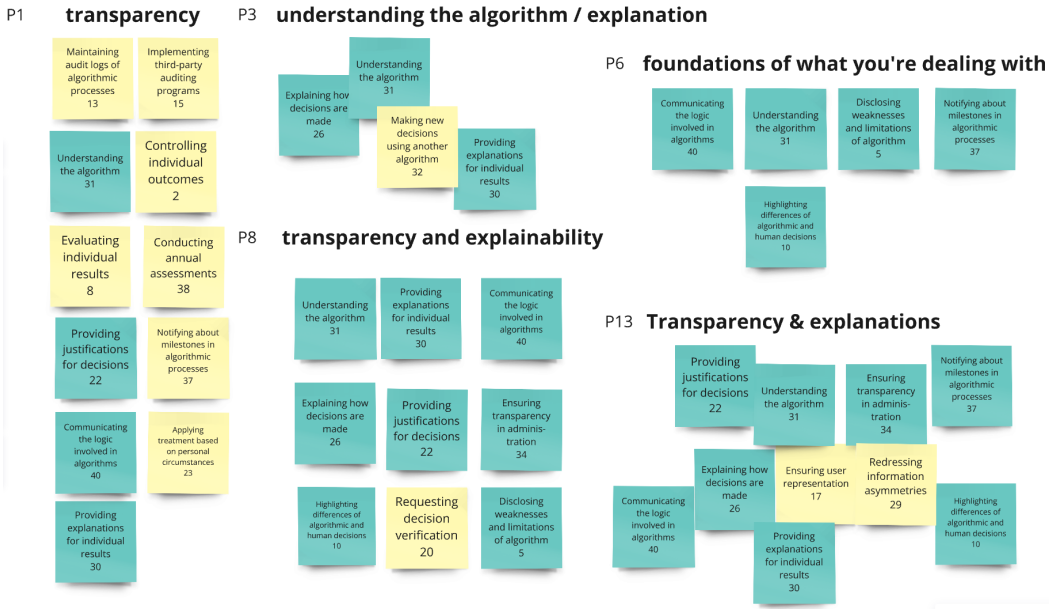

Fig. 3. **Card sort explainability clusters.** Six different card sorts that all contain cards that participants related to transparency or explainability. Often, participants created exactly one cluster with a similar name to those shown. Cards that can be found in multiple of the six clusters are colored petrol, cards that are only in one cluster are colored yellow.

4.2.2 Mechanisms. Realizing the principles of explainability and contestability in actual application contexts requires processes or techniques that implement them. We describe these processes and techniques as *mechanisms*.

In the interviews, participants sorted cards listing explainability and contestability mechanisms into clusters. For explainability, clusters were titled 'transparency', 'understanding the system', 'explanations', or 'foundations of what you're dealing with' (Figure 3). In contrast, contestability appeared less uniform, yielding clusters such as 'control', 'appeals procedure', 'rectification', 'judicial remedies', and 'auditing.' This variety indicates differing understandings across actors, subjects, goals, channels, and implementation modes (Figure 4). Building

on Section 4.1, we map explainability and contestability mechanisms along three axes: whether they provide descriptive or normative information, whether they support individual or collective contestation, and whether they use judicial or non-judicial channels (see Figure 5).

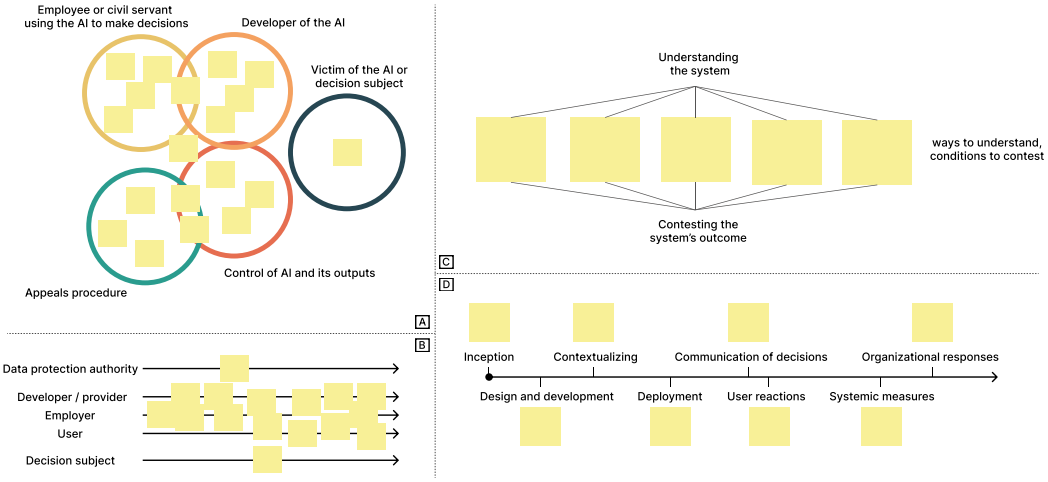

Fig. 4. **Card sort structures.** Participants chose different criteria and dimensions to sort the cards into clusters, including spheres of responsibility (A), responsibility over time and per role (B), ways in which mechanisms connect both to understanding and contesting (C), and an allocation to the implementation process over time (D).

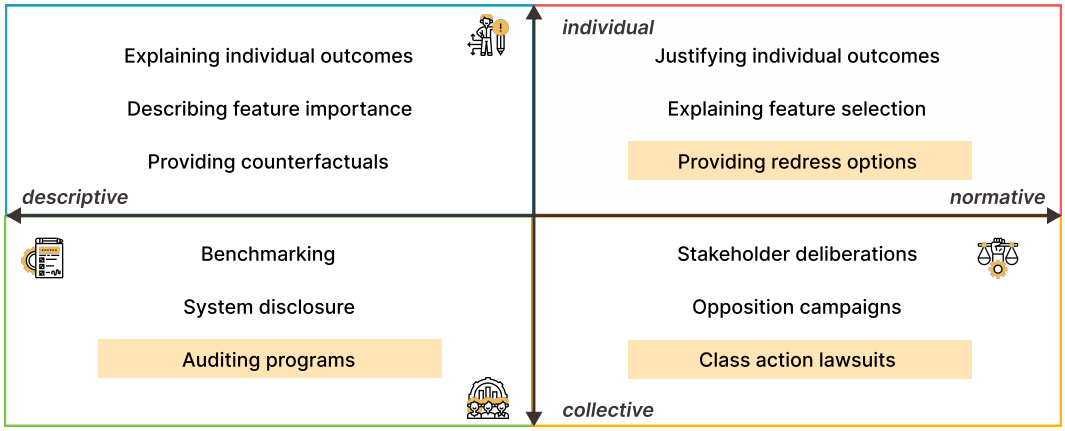

Fig. 5. **Mapping mechanisms of explainability and contestability.** Three dimensions of explainability and contestability: descriptive versus normative approaches (x-axis), individual versus collective action (y-axis), and judicial (highlighted in yellow) versus non-judicial channels. Each quadrant contains examples of mechanisms corresponding to the dimensions. To illustrate, we describe two mechanisms: 'Explaining individual outcomes' means providing decision subjects (individual) with a description of the algorithmic output leading to their decision (descriptive). 'Stakeholder deliberations' can be used to align the values of an AI system with those of the citizens (normative), thus providing a form of democratic control (collective).

4.2.3 Challenges. Explainability and contestability have their limits in alleviating opacity and information asymmetry. Participants articulated three circumstances in which explanation and contestation mechanisms would be ineffective:
i) the system's purpose is to enforce sanctions on decision subjects (i.e., the system is "*badly designed on purpose*" (P4)), ii) the sampling strategy targets vulnerable populations disproportionally, and iii) non-judicial channels of contestation are obscured or not available, i.e.,"*you cannot even exercise [the right to contestation] properly without assistance*" (P9). In consequence, individual explanation and contestation action cannot empower citizens if the policy embedded in an AI system follows values that do not adhere to those of trustworthy AI.

Participants further problematized that individual contestation primarily serves individual interests and has less potential to effect change on a systemic level. Decision subjects "*want to fight for* `their` *privacy and* `their` *freedom. It's very different from fighting for privacy with a big P*" (P12). Collective contestation was posited as a more effective alternative for "*contesting decision patterns [...] through things like class actions*" (P9), but enacting this form of contestation was perceived to be insufficiently explored within the current EU jurisdiction.

Contestability and normative explanations were further perceived as means to create political debate on a collective level. Descriptive explanation do not justify that "*the decision is correct, accurate, and legit*" (P11) and thus lack the normative foundation that is essential for contestability. Participants emphasized the need for justifications, i.e., normative accounts of a system's legitimacy and value alignment (cf. Section 4.1), as a means of providing information that can effectively support contestation.

Lastly, preventing a surge of contestation requests and the gaming of decision processes were perceived as challenges when implementing explanation and contestation measures: "*If [the institutions] do it properly, they will be submerged by requests and contestation*" (P12). Both excessive contestation and gaming attempts may stem from citizens' intention to take control over the decision process to achieve favorable results, especially if they feel that their values are not adequately represented. Participants stressed that when confronted with a badly justified algorithmic decision, citizens "*adapt their behavior to their [antagonistic] interpretation of the algorithm*" (P5). As a remedy, decision subjects could be included in the design process of public AI systems to incorporate their perspectives, improve transparency in rationale, increase acceptance of unfavorable decisions, and fulfill administrative requirements "*to consult the population in terms of impact and gather public feedback*" (P9).

4.3 Gaps between disciplines impede mapping the intersection of explainability and contestability

While the "right to explanation" [21, 103, 110] and judicial ways of contestation, such as appeals and redress, were well known to the interviewed experts, the term "contestability" and the corresponding body of work in design and HCI research were new to some. Participants commented on this, stating "*I'm thinking in which world I was living [...], I didn't notice that [contestability] was so well-developed*" (P3) and "*I'm not familiar with the concept of contestability as such, [...] I rather use 'redress', for example*" (P10). While this points to differences in vocabulary, it also indicates a more conceptual lack of connection between fields. Participants confirmed this gap when exploring the citation graph and stated that the lack of connection between research fields aligned with their experience: "*My experience is, in fact, that the different disciplines are not talking to each other.*" (P8) and "*You have some [communities] that are closer, like [...] legal people sometimes go to the technical part, [...] but some others are not really talking*" (P9). An absence of connection to the legal literature was especially noticed in relation to design, sociotechnical, and ethics literature. Thus, this definitional and conceptual separation between the disciplines, as well as the conceptualization of the relationship between design and regulation, are main challenges in mapping the intersection between explainability and contestability.

5 Explainability and contestability in practice: regulatory, institutional, and technical points of friction

This section is guided by RQ2 and examines points of friction in the implementation of explainability and contestability. We use evidence from the interviews to show that their implementations need to cohere to the spirit of the initial concepts (Section 5.1), that their translation from theory to practice is a shared responsibility among different "regulators" (Section 5.2), and that implementations profit from collaborations between communities (Section 5.3).

5.1 Implementations of explainability and contestability should keep in mind the spirit of the law

Participants identified a key concern of implementing regulatory provisions in the question of whether these implementations would capture the regulation's intent, i.e., the *spirit of the law*. Participants who were members of the HLEG described that the group asserted in unison the importance of explainability in the AI Act: "*lawyers agreed in there, and human rights experts agreed in there, practitioners, [...] there was no doubt that this is a fundamental requirement*" (P8). Explainability was then integrated into the principles and the seven key requirements even though its prospective implementation was already registered as an issue: "*[W]e put as principle something that wasn't really possible 100%. But we felt that it was important to have this because it also reflects on [...] contestability*" (P10). As such, requirements such as those stated in the AI Act are subject to several transformations before being realized in practical applications: they are formalized in legal texts and technical standards, integrated into national jurisdiction, and only then implemented in public institutions. In the course of this process, preserving the spirit of the law is not a given, especially as agreed-upon evaluation metrics for both principles are yet to be developed [44, 80].

In alignment with previous analysis of XAI and law [44], participants feared that downstream applications would not keep the intended safeguards intact due to diverging interpretations. P8 described their experience when meeting lawyers who were "*discussing whether a software application could be a high-risk application. [...] I thought, no, this was not the intention, [it was] to safeguard certain principles*" (P8). Another participant stated that "*it's not just about complying to the letter of the EU AI Act but also the spirit, the spirit of the act is to empower affected individuals to safeguard their rights*" (P11). When understanding design as a vehicle of regulation, then embedding the legislation's intention and corresponding values into design is crucial, rather than designing for compliance "to the letter." We propose one method of value alignment in Section 6.1.

5.2 Implementing explainability and contestability clarifies how responsibility moves between regulators

In the interviews, participants repeatedly considered which actors would be responsible for implementing explanation and contestation mechanisms (these considerations are visually depicted in Figure 4 A and B). Participants considered 'regulators' [90], including policymakers, standardization bodies, data protection authorities, and developers, to be the main actors that shared the "*responsibility to ensure user representation in the development and the use of the AI*" (P5). Technical standards were perceived to be one of the main components to clarify the allocation of responsibility, but their enforcement raised questions. Due to the conceptual polysemy with which legal texts treat both design principles, fulfilling their respective responsibility means that regulators are forced to interpret the provisions, potentially resulting in conflicting points of view between the 'executive' and 'organizational' levels.

Importantly, participants also stated that the regulation of AI systems has not yet actually taken place, as "*nobody has implemented the EU AI Act yet [...]. There's no national law to set down sanctions*" (P6). This has two implications: While first, the task of interpreting the regulatory provisions is not clearly assigned between EU jurisdiction, national authorities, and public institutions, second, the process of assigning responsibility for this

interpretation can still be shaped, leaving room to delineate "*how to handle conflicts*" (P6) and "*how we are going to adapt our legal system*" (P1). We outline ways to create a composite form of responsibility through indirect and direct control mechanisms in Section 6.2.

5.3 Collaboration between communities aids the implementation of explainability and contestability

Participants who had come into contact with both legal and design research on AI regulation regularly highlighted the potential (and current shortcomings) of interdisciplinary collaboration. Participants criticized that technical explanations of AI behavior often were not available "*only because at the beginning of the process, they haven't thought about that*" (P1). In consequence, and because "*explainable AI does not fit into the justification of legal decisions*" (P2), policy-makers were considered to have an incomplete picture of technology. Explainability was described to facilitate communication between disciplines, since "*as soon as we start to explain what we are doing, [...] everyone else, also from different disciplines, can understand*" (P8). In turn, participants proposed that legal studies could improve the normative force of design research by giving it "*a bit more punch*" (P9):

> *[O]ne thing that legal studies can help is to say: 'No, you have to care about this not just because we are a bunch of hippies trying to save the world, but also because if you don't, you're going to have lots of problem with the law [...] or even have your system not being commercialized in a particular jurisdiction.* (P9)

While the benefits of interdisciplinary collaboration are evident, following through was described as "painful" (P8). For those removed from AI's technical side, "*even explaining the concept of explainability sometimes can be challenging because they have to understand that AI is a black box*" (P14). Similar comments concerned non-judicial contestation, which is not well known in legal disciplines (see Section 6.2). Participants with both technical and legal experience found it "very challenging" to "*find a right level of abstraction where we don't get too bogged down on the details [...] versus where we don't generalize too much*" (P9). Our findings are in line with prior legal and conceptual analyses advocating interdisciplinary collaboration to develop shared understandings of explainability [44] and contestability [80]. Future work should unify these approaches and develop methods that foster shared conceptual understanding and vocabulary to inform policy and implementation.

6 Perspectives for AI policy

Drawing from the insights generated in the interviews, we suggest recommendations towards the implementation of explainability and contestability through the lens of Regulation by Design [89] (Section 2.3). To this end, we organize our recommendations in two main groups: (1) those in which design practice contributes to regulation and (2) those in which internal governance systems are used as mechanisms that steer practices in public institutions.

The recommendations are motivated by the public AI system discussed in this study but are not limited to it. The CAF algorithm [95] shares properties with many public-sector AI systems in welfare states [59] that are aimed at efficiency increase through automation, in that it: uses models trained on historical client data [102]; is intended for human review but overlooks pitfalls such as automation bias [45] and algorithmic imprints [34]; and was developed with limited involvement of affected stakeholders and provides few features to enable understanding or contestation [68]. Consequently, the recommendations apply to public AI systems that (i) are used on the public and should thus be publicly deliberated or mandated; (ii) can severely affect individuals and therefore require direct and indirect control mechanisms; and (iii) are tied to administrative procedures that should provide judicial and non-judicial contestation pathways. More generally, we expect them to apply to high-risk AI systems [36] serving the public interest [118].

6.1 Explainability and contestability in regulation by design and by law

We established in Section 2.3 that Regulation by Design views regulation as a rule-making activity performed through social practices like design. As design can simply disable non-compliance, it is a powerful form of regulation. We therefore advocate for assigning greater accountability for design practices that might bypass democratic processes.

6.1.1 Recommendation 1: Making design decisions open to public deliberation. The design of public AI systems involves encoding legislation into software [117]. Design choices on inputs and human–AI interaction become de facto policy that is not publicly deliberated but delegated to third-party developers [82]. To reintroduce deliberation, participants called for stakeholder "early-stage deliberations" [60]. Embedding contestability as a co-designed technical feature, rather than a compliance standard, could align the system values with those of citizens. This may raise decision acceptability while avoiding excessive contestation in operation. We highlight two aspects of participation: i) the participation's level of abstraction, where prior work suggests participatory approaches to focus on values and policies embedded in code, not low-level technical design choices [2, 12]; and ii) the participation mechanisms, where citizen assemblies or advocacy groups can represent those unable to participate (e.g., children) to substitute for direct participation [12]. Yet risks such as gaming and regulatory capture make participation delicate, which we identify as a design–policy tension for future work.

6.2 Strengthening the intersection of explainability and contestability through internal AI governance

Internal governance systems coordinate social action through formal and informal mechanisms. Governance acts as a meta-regulative activity steering how other practices, like design, regulate. We suggest that internal AI governance systems strengthens the implementation of explainability and (non-judicial) contestability mechanisms.

6.2.1 Recommendation 2: Combining indirect and direct control mechanisms. In the interviews, a consensus emerged that contestability lets individuals exercise control over public-sector AI, based on explainability. Individual action was seen as suitable for assessing specific decisions' acceptability; collective action to judging a system's overall legitimacy. Legal research favors combining indirect control (a regulator) with direct oversight by decision subjects who can appeal and obtain redress ("democratic control" [12]) [47, 105, 108]. Using contestability to give subjects direct oversight was viewed positively by AI regulation experts but they were unfamiliar with concrete implementations. The balance between indirect and direct control should be revisited, e.g., strengthening direct control via non-judicial contestation. This is aided by explanations that disclose a system's purpose (normative) rather than merely its workings (descriptive) [12]. Institutions should consider what forms of contestation explanations enable [103] and ensure they are relevant and intelligible [24, 85]. They can also engage regulators (e.g., AI Office) and standardizers for support in identifying direct and indirect mechanisms when implementing explainability and contestability, and aligning with regulatory intent [90]. Tools such as the Contestability Assessment Score [80] can help operationalize dimensions discussed here (e.g., audits, access to contestation). However, integrating our mapping of mechanisms to individual vs. collective and descriptive vs. normative dimensions into such operationalizations remains future work.

6.2.2 Recommendation 3: Ensuring the availability of non-judicial contestation channels. The majority of the interviewees were familiar with judicial means of contestation and unfamiliar with non-judicial ones, which might explain why this aspect of contestability remained overlooked in regulatory initiatives until now [57]. We argue that public institutions should adopt a more holistic approach to contestability that goes beyond "complying to the letter" to improve trust and acceptability. To this end, non-judicial means to implement contestability could be better leveraged in legal instruments guiding the implementation of AI regulations, especially as standards

are developing into means of judicial control [46, 99]. While policy needs to decide *what* can be contested, *who* can contest, *who* is accountable, and *which types of review* should be put in place [71], internal AI governance systems should ensure ways for non-judicial contestation. These include tools for scrutiny, annual assessments, or differential treatment (i.e., room to negotiate decisions between decision subjects and operators) [5]. Such mechanisms should ensure that decision subjects are given an opportunity to understand their situation [73] and can articulate their voice in the process [116]. To design and implement these mechanisms alongside AI systems, public institutions can benefit from engaging with design and HCI researchers. This form of "system-people-policy interaction" [113] presents a difficult but fruitful avenue for future work on explainability and contestability in the intersection between research and policy.

7 Limitations

Like any research, this study had limitations. *First*, the single 30-minute card sort may have limited exploration and classification depth, and the citation graph may have overwhelmed participants unfamiliar with such representations. However, because the analysis focused on participants' interdisciplinary experiences and perceptions rather than detailed graph analysis, design and interaction effects should be minimal. *Second*, the study focused on the CAF algorithm and logistic regression due to their widespread public-sector use. Future research could examine how the linkage between explainability and contestability aligns or differs in other application contexts and compare this to the empirical insights here. *Third*, we reached data saturation (no new codes) after interviewing 14 participants from European universities and public institutions. Saturation dynamics might have differed with experts beyond Europe. *Fourth*, the open-ended task enabled contextual exploration and adaptation to participants' expertise. A more structured approach linking the two principles (e.g., predefined relationships) could identify specific properties but might limit the broader meanings central to our findings. While our format suited a cross-domain sample, a more structured design may better support analysis within a clearly defined ADM context. *Fifth*, our query erroneously excluded "right to explanation"; nonetheless, relevant papers likely use alternative terms included in our query (e.g., "artificial intelligence," "right to," "explain*"), so relevant literature should still be captured. As our findings draw on participants' perceptions and reflect general trends rather than exhaustive coverage, conclusions about citation dynamics and interpretive patterns remain valid. *Finally*, the analysis excluded jurisdiction comparisons, a fruitful direction for future research.

8 Conclusion

In this work, we conceptualized explainability and contestability and their implementation by interviewing 14 interdisciplinary experts on AI regulation. We provided distinctions to facilitate implementation of these principles, including between normative and descriptive explanations, individual and collective contestation action, and judicial and non-judicial contestation channels. Key challenges include the preservation of the regulation's spirit, the responsibility for interpreting regulatory provisions, and the collaboration between disciplines. Based on these findings, we recommend i) strengthening the intersection between both principles in policy and governance, ii) considering non-judicial channels of contestation to improve trust, and iii) employing early-stage deliberations in the development of public AI systems to improve acceptability and avoid excessive contestation. With these insights, we aim to inform research and policy efforts that leverage explainability and contestability in the development of trustworthy public AI systems.

Acknowledgments

This work has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20058] as well as [10.47379/ICT20065]. This research was further funded through the French ANR project CHAI (ANR-24-CE38-6380), and was made possible by the support of the French Embassy in Austria.

References

[1] 2024. France: Discriminatory algorithm used by the social security agency must be stopped — amnesty.org. https://www.amnesty.org/en/latest/news/2024/10/france-discriminatory-algorithm-used-by-the-social-security-agency-must-be-stopped/.

[2] Amina A. Abdu, Lauren M. Chambers, Deirdre K. Mulligan, and Abigail Z. Jacobs. 2024. Algorithmic Transparency and Participation through the Handoff Lens: Lessons Learned from the U.S. Census Bureau's Adoption of Differential Privacy. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1150–1162. doi:10.1145/3630106.3658962

[3] Andrea Aler Tubella, Andreas Theodorou, Virginia Dignum, and Loizos Michael. 2020. Contestable Black Boxes. In *Rules and Reasoning*, Víctor Gutiérrez-Basulto, Tomáš Kliegr, Ahmet Soylu, Martin Giese, and Dumitru Roman (Eds.). Springer International Publishing, Cham, 159–167. doi:10/gjd5pg

[4] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by Design: Towards a Framework. *Minds and Machines* 33, 4 (01 12 2023), 613–639. doi:10.1007/s11023-022-09611-z

[5] Kars Alfrink, Ianus Keller, Mireia Yurrita Semperena, Denis Bulygin, Gerd Kortuem, and Neelke Doorn. 2024. Envisioning Contestability Loops: Evaluating the Agonistic Arena as a Generative Metaphor for Public AI. *She Ji: The Journal of Design, Economics, and Innovation* 10, 1 (2024), 53–93.

[6] Marco Almada. 2019. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. ACM, Montreal QC Canada, 2–11. doi:10.1145/3322640.3326699

[7] and European Economic and Social Committee. 2021. *Digital Services Act and Digital Markets Act – Stepping stones to a level playing field in Europe*. European Economic and Social Committee. doi:doi/10.2864/28842

[8] Ricardo Baeza-Yate and Jeanna Matthews. 2022. Statement on Principles for Responsible Algorithmic Systems. https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf Last accessed on 13th May 2024.

[9] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1194–1206. doi:10.1145/3531146.3533179

[10] Julia Black. 2001. Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a 'Post-Regulatory' World. *Current Legal Problems* 54, 1 (12 2001), 103–146. arXiv:https://academic.oup.com/clp/article-pdf/54/1/103/7524076/54-1-103.pdf doi:10.1093/clp/54.1.103

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (oct 2008), P10008. doi:10.1088/1742-5468/2008/10/P10008

[12] Daniel James Bogiatzis-Gibbons. 2024. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 74–84. doi:10.1145/3630106.3658541

[13] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike Von Luxburg. 2022. Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 891–905. doi:10.1145/3531146.3533153

[14] Marco Borraccetti. 2011. *Fair Trial, Due Process and Rights of Defence in the EU Legal Order*. 95–107. doi:10.1007/978-94-007-0156-4_5

[15] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (<conf-loc>, <city>Helsinki</city>, <country>Finland</country>, </conf-loc>) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 807–819. doi:10.1145/3490099.3511139

[16] Clara Bove, Thibault Laugel, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2024. Why do explanations fail? A typology and discussion on failures in XAI. http://arxiv.org/abs/2405.13474

[17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[18] Virginia Braun and Victoria Clarke. 2023. Approaches to thematic analysis: Becoming a knowing researcher. *Starting research in clinical education* (2023), 165–174.

[19] Stephen G. Breyer. 2009. *Regulation and Its Reform*. Harvard University Press, Cambridge.

[20] Fred Carden. 2009. *Knowledge to policy: making the most of development research*. SAGE, Los Angeles, Calif.

[21] Casey, Bryan; Farhangi, Ashkon; Vogl, Roland. 2019. Rethinking Explainable Machines: The GDPR's Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise. (2019). doi:10.15779/Z38M32N986 Publisher: btlj.

[22] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. doi:10.1089/big.2016.0047

[23] Danielle Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Washington Law Review* 89 (March 2014), 1–33.

[24] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 598–609.

[25] Mark Coeckelbergh. 2020. *AI ethics*. The MIT Press.

[26] Council of Europe. 2018. *Convention 108+: Convention for the protection of individuals with regard to the processing of personal data*. https://rm.coe.int/convention-108-convention-for-the-protection-of-individuals-with-regar/16808b36f1

[27] Court of Justice of the European Union. 2023. Case C-634/21, SCHUFA Holding (Scoring). ECLI:EU:C:2023:957.

[28] Court of Justice of the European Union. 2023. Joined Cases C-26/22 and C-64/22, SCHUFA Holding. ECLI:EU:C:2023:958.

[29] Court of Justice of the European Union. 2024. Case C-203/22, Dun & Bradstreet Austria. ECLI:EU:C:2025:117.

[30] Rebecca Crootof, Margot E Kaminski, and W Nicholson Price Ii. 2023. Humans in the Loop. *VANDERBILT LAW REVIEW* 76 (2023). https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=2659&context=law-faculty-publications

[31] Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. A Critical Survey on Fairness Benefits of Explainable AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1579–1595. doi:10.1145/3630106.3658990

[32] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. ACM, Virtual Event USA, 1591–1602. doi:10.1145/3461778.3462131

[33] Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review* 16, 1 (Dec. 2017), 18–84.

[34] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The Algorithmic Imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1305–1317. doi:10.1145/3531146.3533186

[35] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc.

[36] European Commission. 2024. Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations. .

[37] European Parliament and Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. https://data.europa.eu/eli/reg/2016/679/oj

[38] European Parliament and Council of the European Union. 2021. Proposal for a directive of the European Parliament and the Council on improving working conditions in platform work. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0762

[39] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1457–1466. doi:10.1145/3531146.3533202

[40] Clàudia Figueras, Harko Verhagen, and Teresa Cerratto Pargman. 2021. Trustworthy AI for the People?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 269–270. doi:10.1145/3461702.3462470

[41] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (June 2019), 261–262. doi:10.1038/s42256-019-0055-y

[42] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28, 4 (Dec. 2018), 689–707. doi:10.1007/s11023-018-9482-5

[43] Timo Freiesleben and Gunnar König. 2023. Dear XAI Community, We Need to Talk!. In *Explainable Artificial Intelligence*, Luca Longo (Ed.). Springer Nature Switzerland, Cham, 48–65. doi:10.1007/978-3-031-44064-9_3

[44] Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco F. Huber, and Christian Horz. 2024. How Should AI Decisions Be Explained? Requirements for Explanations from the Perspective of European Law. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 438–450. doi:10.1609/aies.v7i1.31648

[45] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2011. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (06 2011), 121–127. doi:10.1136/amiajnl-2011-000089

[46] Mélanie Gornet and Winston Maxwell. 2024. The European approach to regulating AI through technical standards. *Internet Policy Review* 13, 3 (July 2024). doi:10.14763/2024.3.1784

[47] John Graham, Timothy Wynne Plumptre, and Bruce Amos. 2003. *Principles for good governance in the 21st century*. Vol. 15. Institute on governance Ottawa.

[48] Balint Gyevnar, Nick Ferguson, and Burkhard Schafer. 2023. Bridging the Transparency Gap: What Can Explainable AI Learn from the AI Act? In *Frontiers in Artificial Intelligence and Applications*, Kobi Gal, Ann Nowé, Grzegorz J. Nalepa, Roy Fairstein, and Roxana Rădulescu (Eds.). IOS Press. doi:10.3233/FAIA230367

[49] Clément Henin and Daniel Le Métayer. 2022. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* 37, 4 (Dec. 2022), 1397–1410. doi:10.1007/s00146-021-01251-8

[50] Monique M. Hennink, Bonnie N. Kaiser, and Vincent C. Marconi. 2017. Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research* 27, 4 (March 2017), 591–608. doi:10.1177/1049732316665344

[51] High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. (April 2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-{AI}

[52] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Z. E. Imel, and D. C. David C. Atkins. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. In *DIS 2017 - Proceedings of the 2017 ACM Conference on Designing Interactive Systems*. ACM Press, 95–99. doi:10/gddxqb

[53] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. 2023. Explainable AI: roles and stakeholders, desirements and challenges. *Frontiers in Computer Science* 5 (Aug. 2023), 1117848. doi:10.3389/fcomp.2023.1117848

[54] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1395–1417. doi:10.1145/3630106.3658979

[55] Isabelle Hupont, Marina Micheli, Blagoj Delipetrev, Emilia Gómez, and Josep Soler Garrido. 2023. Documenting High-Risk AI: A European Regulatory Perspective. *Computer* 56, 5 (May 2023), 18–27. doi:10.1109/MC.2023.3235712 Conference Name: Computer.

[56] Steven J. Jackson, Tarleton Gillespie, and Sandy Payette. 2014. The Policy Knot: Re-Integrating Policy, Practice and Design in Cscw Studies of Social Computing. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 588–602. doi:10/gg5g9w

[57] Margot E Kaminski and Jennifer M Urban. 2021. The right to contest AI. *Columbia Law Review* 121, 7 (2021), 1957–2048.

[58] Margot E Kaminski and Jennifer M Urban. 2021. The Right to Contest AI. *Columbia Law Review* 121, 7 (Nov. 2021), 92.

[59] Anne Kaun. 2022. Suing the algorithm: the mundanization of automated decision-making in public services through litigation. *Information, Communication & Society* 25, 14 (Oct. 2022), 2046–2062. doi:10.1080/1369118X.2021.1924827

[60] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 749, 22 pages. doi:10.1145/3613904.3642849

[61] Caitlin Kearney, Jiri Hron, Helen Kosc, and Miri Zilka. 2024. Beyond Use-Cases: A Participatory Approach to Envisioning Data Science in Law Enforcement. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1809–1826. doi:10.1145/3630106.3659007

[62] Perry Keller and Archie Drake. 2021. Exclusivity and Paternalism in the Public Governance of Explainable AI. *Computer Law & Security Review* 40 (April 2021), 105490. doi:10/g9rjwt

[63] Klára Kolářová and Timothée Schmude. 2026. Start Using Justifications When Explaining AI Systems to Decision Subjects. In *Digital Humanism*, Ludger Hagedorn, Ute Schmid, Susan Winter, and Stefan Woltran (Eds.). Springer Nature Switzerland, Cham, 190–202.

[64] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (July 2021), 103473. doi:10.1016/j.artint.2021.103473

[65] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[66] Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, Xiang Yin, Dekai Zhang, and Francesca Toni. 2024. Contestable AI Needs Computational Argumentation. *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning* 21, 1 (Oct. 2024), 888–896. doi:10/g82hjc

[67] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Commun. ACM* 61, 10 (Sept. 2018), 36–43. doi:10.1145/3233231

[68] Paola Lopez. 2019. Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. In *Proceedings of the 18th Annual STS Conference*. Graz, 289–309. doi:10.3217/978-3-85125-668-0-16

[69] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[70] Henrietta Lyons, Tim Miller, and Eduardo Velloso. 2023. Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 764–774. doi:10.1145/3593013.3594041

[71] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 106 (April 2021), 25 pages. doi:10.1145/3449180

[72] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Designing for Contestation: Insights from Administrative Law. http://arxiv.org/abs/2102.04559 arXiv:2102.04559 [cs].

[73] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. doi:10.1145/3491102.3517606

[74] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. doi:10.1145/3313831.3376445

[75] David Mair, Laura Smillie, Giovanni La Placa, Florian Schwendinger, Milena Raykovska, Zsuzsanna Pasztor, René van Bavel, and European Commission (Eds.). 2019. *Understanding our political nature: how to put knowledge and reason at the heart of political decision-making*. Publications Office, Luxembourg. doi:10.2760/374191

[76] Joseph A. Maxwell. 2010. Using Numbers in Qualitative Research. *Qualitative Inquiry* 16, 6 (2010), 475–482. doi:10.1177/1077800410364740

[77] Winston Maxwell and Bruno Dumas. 2023. Meaningful XAI based on user-centric design methodology: Combining legal and human-computer interaction (HCI) approaches to achieve meaningful algorithmic explainability. *SSRN Electronic Journal* (2023). doi:10.2139/ssrn.4520754

[78] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. doi:10.1016/j.artint.2018.07.007

[79] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). https://christophm.github.io/interpretable-ml-book

[80] Catarina Moreira, Anna Palatkina, Dacia Braca, Dylan M. Walsh, Peter J. Leihn, Fang Chen, and Nina C. Hubig. 2025. Explainable AI Systems Must Be Contestable: Here's How to Make It Happen. arXiv:2506.01662 [cs.CY] https://arxiv.org/abs/2506.01662

[81] Bronwen Morgan and Karen Yeung. 2007. *An Introduction to Law and Regulation: Text and Materials*. Cambridge University Press.

[82] Deirdre K Mulligan and Kenneth A Bamberger. 2019. Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ* 34 (2019), 773.

[83] Nadia Nahar, Jenny Rowlett, Matthew Bray, Zahra Abba Omar, Xenophon Papademetris, Alka Menon, and Christian Kästner. 2024. Regulating Explainability in Machine Learning Applications – Observations from a Policy Design Experiment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2101–2112. doi:10.1145/3630106.3659028

[84] Nataliya Nedzhvetskaya and JS Tan. 2024. No Simple Fix: How AI Harms Reflect Power and Jurisdiction in the Workplace. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 422–432. doi:10.1145/3630106.3658915

[85] Chris Norval, Kristin Cornelius, Jennifer Cobbe, and Jatinder Singh. 2022. Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 679–690.

[86] OECD. 2019. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[87] P Jonathon Phillips, Carina A Hahn, Peter C Fontana, Amy N Yates, Kristen Greene, David A Broniatowski, and Mark A Przybocki. 2021. *Four principles of explainable artificial intelligence*. Technical Report NIST IR 8312. National Institute of Standards and Technology (U.S.), Gaithersburg, MD. NIST IR 8312 pages. doi:10.6028/NIST.IR.8312

[88] Yulu Pi and Maddie Proctor. 2025. Toward empowering AI governance with redress mechanisms. *Cambridge Forum on AI: Law and Governance* 1 (2025). doi:10.1017/cfl.2025.9 Publisher: Cambridge University Press (CUP).

[89] Kostina Prifti. 2024. The Theory of 'Regulation By Design' : Towards a Pragmatist Reconstruction. *Technology and Regulation* 2024 (Aug. 2024), 152–166. doi:10/g9dr24

[90] Kostina Prifti, Jessica Morley, Claudio Novelli, and Luciano Floridi. 2024. Regulation by Design: Features, Practices, Limitations, and Governance Implications. *Minds and Machines* 34, 2 (May 2024), 13. doi:10.1007/s11023-024-09675-z

[91] Carina Prunkl. 2022. Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence* 4, 2 (Feb. 2022), 99–101. doi:10.1038/s42256-022-00449-9

[92] Muhammad Raees, Inge Meijerink, Ioanna Lykourentzou, Vassilis-Javed Khan, and Konstantinos Papangelis. 2024. From Explainable to Interactive AI: A Literature Review on Current Trends in Human-AI Interaction. *International Journal of Human-Computer Studies* 189 (Sept. 2024), 103301. doi:10/gtwr25

[93] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 959–972. doi:10.1145/3531146.3533158

[94] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[95] Manon Romain, Adrien Senecat, Soizic Pénicaud, Gabriel Geiger, and Justin-Casimir Braun. 2023. How We Investigated France's Mass Profiling Machine — lighthousereports.com. https://www.lighthousereports.com/methodology/how-we-investigated-frances-mass-profiling-machine/.

[96] Manon Romain, Adrien Sénécat, Elsa Delmas, Thomas Steffen, Léa Girardot, and Lighthouse Reports. 2024. Comment l'algorithme de la CAF prédit si vous êtes « à risque» de frauder — lemonde.fr. https://www.lemonde.fr/les-decodeurs/visuel/2023/12/04/comment-l-algorithme-de-la-caf-predit-si-vous-etes-a-risque-de-frauder_6203836_4355770.html.

[97] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (01 5 2019), 206–215. doi:10.1038/s42256-019-0048-x

[98] Gordon Rugg and Peter McGeorge. 2005. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems* 22, 3 (2005), 94–107.

[99] Harm Schapel. 2013. The New Approach to the New Approach: The Juridification of Harmonized Standards in EU Law. *Maastricht Journal of European and Comparative Law* 20, 4 (2013), 521–533. doi:10.1177/1023263X1302000404

[100] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2024. Information That Matters: Exploring Information Needs of People Affected by Algorithmic Decisions. arXiv:2401.13324 [cs.HC]

[101] David Schneeberger, Richard Röttger, Federico Cabitza, Andrea Campagner, Markus Plass, Heimo Müller, and Andreas Holzinger. 2023. The Tower of Babel in Explainable Artificial Intelligence (XAI). In *Machine Learning and Knowledge Extraction*, Andreas Holzinger, Peter Kieseberg, Federico Cabitza, Andrea Campagner, A Min Tjoa, and Edgar Weippl (Eds.). Vol. 14065. Springer Nature Switzerland, Cham, 65–81. doi:10.1007/978-3-031-40837-3_5

[102] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, Seoul Republic of Korea, 2138–2148. doi:10.1145/3531146.3534631

[103] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (Nov. 2017), 233–242. doi:10.1093/idpl/ipx022

[104] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–29. doi:10.1145/3479577

[105] Nathalie A Smuha. 2021. Beyond the individual: governing AI's societal harm. *Internet Policy Review* 10, 3 (2021).

[106] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy Artificial Intelligence. *Electronic Markets* 31, 2 (June 2021), 447–464. doi:10.1007/s12525-020-00441-4

[107] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–28. doi:10.1145/3476059

[108] Sandra Wachter. 2024. Limitations and loopholes in the EU AI Act and AI Liability Directives: what this means for the European Union, the United States, and beyond. *Yale Journal of Law and Technology* 26, 3 (2024).

[109] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (May 2017), 76–99. doi:10.1093/idpl/ipx005

[110] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal* (2017). doi:10.2139/ssrn.3063289

[111] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. doi:10.1145/3397481.3450650

[112] Georg Wenzelburger, Pascal D. König, Julia Felfeli, and Anja Achtziger. 2022. Algorithms in the public sector. Why context matters. *Public Administration* (Dec. 2022), padm.12901. doi:10.1111/padm.12901

[113] Qian Yang, Richmond Y. Wong, Steven Jackson, Sabine Junginger, Margaret D. Hagan, Thomas Gilbert, and John Zimmerman. 2024. The Future of HCI-Policy Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3613904.3642771

[114] Karen Yeung. 2015. Design for the Value of Regulation. In *Handbook of Ethics, Values, and Technological Design*, Jeroen Van Den Hoven, Pieter E. Vermaas, and Ibo Van De Poel (Eds.). Springer Netherlands, Dordrecht, 447–472. doi:10.1007/978-94-007-6970-0_32

[115] Mireia Yurrita, Agathe Balayn, and Ujwal Gadiraju. 2023. Generating Process-Centric Explanations to Enable Contestability in Algorithmic Decision-Making: Challenges and Opportunities. In *2023 Human-Centered XAI Workshop at CHI Conference on Human Factors in Computing Systems (CHI '23)*.

[116] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–21. doi:10.1145/3544548.3581161

[117] Stavros Zouridis, Marlies van Eck, and Mark Bovens. 2020. *Automated Discretion*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-19566-3_20

[118] Theresa Züger and Hadi Asghari. 2023. AI for the public. How public interest theory shifts the discourse on AI. *AI & SOCIETY* 38, 2 (April 2023), 815–828. doi:10.1007/s00146-022-01480-5

9 Endmatter

9.1 Generative AI Disclosure Statement

During the preparation of this manuscript, we utilized the following AI-assisted tool to assist with formatting, grammar, and fluency of writing: ChatGPT (version GPT-5, released by OpenAI in 2025), Claude (version Opus 4.5 released by Anthropic in 2025), Writefull, and Grammarly. These tools were used solely to improve clarity and readability without altering the paper's intellectual content, methodology, or findings.

10 Use and attribution of icons

The icons "post-it", "choice", "decision-making", "direction", "cursor" used in Figure 2 are provided by Freepik, the icon "knowledge graph" used in the same figure is provided by Grafixpoint through Flaticon.com. The icons used in Figure 5 are provided by Eucalyp through Flaticon.com.

Appendix

A Screening criteria of participants

All study participants were screened for a range of criteria, which included:

- Field of education and highest diploma (free text)
- Current occupation (free text)
- Experience in current field of work (Less than 2 years, between 2 and 5 years, between 6 and 10 years, more than 10 years)
- Self-rated knowledge of technical aspects of AI systems, such as machine learning algorithms and data processing (4+1-point scale: not knowledgeable, slightly knowledgeable, somewhat knowledgeable, very knowledgeable; not sure)
- Self-rated knowledge of current regulations and policies regarding use and deployment of AI systems (4+1-point scale: not knowledgeable, slightly knowledgeable, somewhat knowledgeable, very knowledgeable; not sure)

Table 1 lists participants' responses with respect to their experience in the field of work and their knowledge regarding technical and regulatory aspects. The occupation was omitted to retain anonymity of participants.

ID	Experience	Technical AI Knowledge	Regulation Knowledge
1	More than 10 years	Very knowledgeable	Very knowledgeable
2	More than 10 years	Somewhat knowledgeable	Very knowledgeable
3	More than 10 years	Slightly knowledgeable	Very knowledgeable
4	Between 6 and 10 years	Somewhat knowledgeable	Very knowledgeable
5	More than 10 years	Somewhat knowledgeable	Very knowledgeable
6	Between 2 and 5 years	Somewhat knowledgeable	Very knowledgeable
7	More than 10 years	Very knowledgeable	Somewhat knowledgeable
8	More than 10 years	Somewhat knowledgeable	Very knowledgeable
9	Between 6 and 10 years	Somewhat knowledgeable	Very knowledgeable
10	More than 10 years	Very knowledgeable	Very knowledgeable
11	Between 2 and 5 years	Very knowledgeable	Very knowledgeable
12	More than 10 years	Somewhat knowledgeable	Somewhat knowledgeable
13	Between 2 and 5 years	Somewhat knowledgeable	Slightly knowledgeable
14	Less than 1 year	Very knowledgeable	Somewhat knowledgeable

Table 1. Participant experience and self-assessed knowledge of AI technology and regulations.

B Interview guide

The following describes the study procedure, elements, and questions that guided the semi-structured interviews with participants.

(1) Give consent form and questionnaire
(2) Introduction to study
   a. The study is about the relation of explainability and contestability, how one supports the other, and your theoretical and practical experience in dealing with these topics.
   b. The process is composed of three parts: card sort, use case discussion, and citation graph exploration.
   c. To start, could you please briefly describe your work and role in your institution? Can you tell us about the types of projects or tasks you typically work on in your role?

d. Could you briefly walk us through what you think about when you think about an explainable process? And a contestable one?

(3) Card sort: List of explanation and contestation mechanisms
   a. Prompt: Please sort the cards into groups that make sense to you and think aloud.
   b. You can make yourself familiar with all cards before beginning the sort.
   c. If something is missing, you can add new cards.
   d. At the end, each group is named and shortly discussed.
   e. Feel free to ask questions!
   f. (After sorting) How is each cluster called? Why are these cards in there?
   g. For unobservable/subjective categories: How can you tell? What does this mean?

(4) Use case discussion: Social security risk scoring algorithm
   a. What are your first thoughts about this case? Does it sound familiar?
   b. Taking this example, how would you advise Juliette to obtain information about the decision-making process? And to take action to contest future decisions?
   c. If you could now add other mechanisms for explanation and contestation that would be better suited to this case, which would they be?
   d. The agency asks for your advice on implementing new explanation and contestation mechanisms. How would you advise them to proceed?
   e. Are there any mechanisms or procedural steps of explainability and contestability that are not required by law in such a case, but which should be required?

(5) Graph discussion: Citation and semantic graph.
   a. What are your first thoughts when looking at this graph?
   b. Where would you situate your own work in this graph?
   c. Do the connections between clusters correspond to your experience, or not?
   d. How are explainability and contestability connected here from your perspective?
   e. Is a connection missing that should be there? Why?
   f. How would you label these clusters? Would you label them differently?
   g. If someone had all resources and means, which kind of research should they conduct to give more insight on the topics discussed, explainability and contestability?

(6) Closing
   a. Do you have any more thoughts or comments?
   b. Does anyone come to mind whom you could recommend as an interviewee?

C Card sort material

During the study, participants were presented with a card sort task to elicit their mental models of explanation and contestaitn mechanisms and concepts. Figure 6 depicts all cards used in this task.

The cards were compiled from literature on explainability and contestability and included elements that previous work identified as a relevant concept or mechanism. The resulting list of elements was thematically analyzed to identify common elements, cleaned of duplicates, and formatted to fit the concept cards. The set was then tested in two pilot studies to clear up formulation issues and unclarities. The content of the cards was intended to be grounded in current research while allowing for a flexible interpretation by participants. Table 2 lists the literature sources and lists of mechanisms and concepts that were used to create the cards.

Fig. 6. **Card sort material.** The image depicts the full selection of 40 cards with explainability and contestability mechanisms. Participants received these cards and were asked to sort them into self-defined categories. Numbers were assigned at random, serving as IDs. New items could be added using the stack of empty cards.

Table 2. Concepts and mechanisms for explainability and contestability in AI systems across prior literature.

Source	Concepts and mechanisms
Alfrink et al. [5]	Annual assessments Differential treatment Explanations Input data revisions Interactive controls Intervention requests Monitoring Participatory policy-making Participatory system development Pro-active notifications Pro-social behaviour incentives Tools for scrutiny
Citron and Pasquale [23]	Audit trails of the correlations and inferences made by the algorithm Disclosure of algorithm to neutral experts Interactive controls for customers
Lee et al. [65]	Allow end-users to control the input Appeal or modify decision after it has been made Communicate decision criteria Communicate rules and methods of decision-making Communicate values and concerns of decision members Demonstrate difference between human and algorithmic decisions Elect users to be spokespeople for other users

Continued on next page

Source	Concepts and mechanisms
	Explain how the decision came to be
	Let users control group outcome through discussion and alternatives
	Let users control their individual outcome
	Let users provide input and feedback
	Show difficulties and limitations of ADM
Lyons et al. [72]	Auditing program for third parties
	Fair hearing if decision is adverse
	Internal merits review, external merits review, judicial review
	Limit contestation to people who are affected
	Notify about ways to initiate a review
	Statement of reasons containing evidence and reasons for decision
Lyons et al. [73]	Provide voice to the user
Lyons et al. [70]	Provide new contextual information
	Request new decision
	Request verification
Maxwell and Dumas [77]	Communicate the logic involved
	Control over algorithmic performance
	Disclose weaknesses and use limitations
	Empowerment of users
	Enhance public administration transparency
	Evaluation of individual results
	Local explanations of individual decisions
	Permit users to understand algorithmic decisions and adapt to them
	Provide reasons underpinning algorithmic decisions
	Redressing information asymmetries
Shen et al. [104]	Everyday audits by users
Vaccaro et al. [107]	Compassion in algorithmic decisions
	Representation of user
	Two-way communication between user and provider
Wachter et al. [110]	Auditing APIs
	Human assessment with algorithmic elements
	Human makes decision without algorithmic help
	Human monitors data and processing of algorithmic decision
	Let users express their view
	New decision is made by algorithm
Yurrita et al. [115]	Let user influence decision outcome

D Citation graph

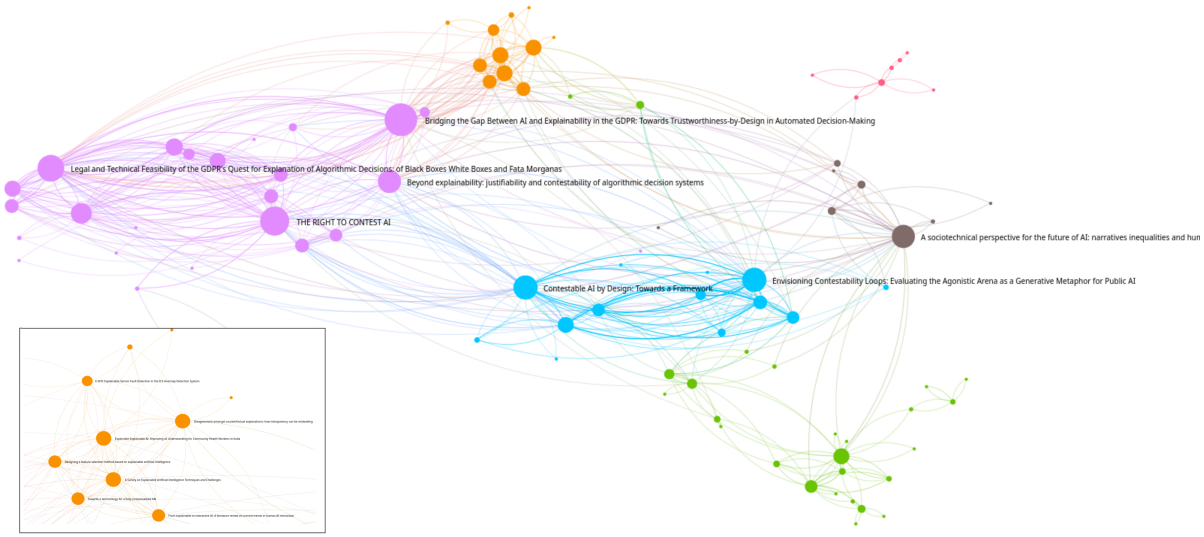

Fig. 7. **Overview of the citation graph.** The image shows the citation graph used in the study to elicit participants' reflections on the research landscape. The graph is the largest connected component of the co-reference graph related to contestability and AI, and it also includes references on the 'right to explanation.' The detail shows a zoom on the "explainable AI" cluster in the network.

E Themes and codes from the inductive thematic analysis

Table 3. This code book contains codes and themes, descriptions, and code counts from the thematic analysis of the interviews.

Themes: Codes	Description	Count
Contestability	General references to contesting AI decisions and systems	200
Contestability: challenges	Obstacles and barriers to effective contestation	20
Contestability: compassion	The role of empathy and care in contestation processes	13
Contestability: contestation that is collective	Collective or group-based forms of contestation	14
Contestability: contestation that is individualized	Individual-level approaches to contestation	24
Contestability: definition	Attempts to define contestability and its scope	12
Contestability: elements	Core components or building blocks of contestation	32
Contestability: Goals of contestation	Intended outcomes and aims of contestation	8
Contestability: Judicial vs non-judicial contestation	Comparison of legal and other contestation mechanisms	2
Contestability: power dynamics	How power relations shape contestation	12
Contestability: Problematic contestation	Cases where contestation is counterproductive or harmful	7
Contestability: suggestions	Proposed improvements or practices for contestation	32
Contestability: unclarities	Ambiguities or unclear aspects of contestability	24
Explainability	General references to making AI understandable	149
Explainability: acceptability	How explanations influence acceptance of algorithmic decisions	9
Explainability: AI literacy	The role of user education and literacy in understanding AI	15
Explainability: challenges	Difficulties in creating useful, accurate explanations	19
Explainability: definition	Attempts to define explainability and its scope	27
Explainability: elements	Key features and components of explanations	39
Explainability: folk theory	Informal mental models people use to interpret AI	2
Explainability: public AI narratives	Broader societal stories and narratives about AI	7
Explainability: purposes	The reasons explanations are provided (e.g., trust, accountability)	7
Explainability: suggestions	Proposed methods and practices to improve explainability	24
Institutional implementation	Organizational practices and structures for operationalizing AI governance	71
Institutional implementation: Institutional landscape	Mapping institutions, roles, and authority lines	27
Institutional implementation: Internal governance	Internal policies, oversight, and risk management	26
Institutional implementation: Ombudsman oversight	External or independent complaint-handling and oversight	8
Institutional implementation: Specification	Translating principles into technical and procedural specifications	2
Institutional implementation: Trade-offs	Balancing competing values and constraints in institutions	8
Intersection Explainability - Contestability	How explainability and contestability interact and reinforce each other	117
Intersection Explainability - Contestability: Connecting explainability and contestability	Mechanisms linking explanations to meaningful contestation	33
Intersection Explainability - Contestability: Justification	Using explanations to justify decisions in contestation contexts	12
Intersection Explainability - Contestability: Limits of explainability and contestability	Boundaries and constraints of both practices	26
Intersection Explainability - Contestability: Making sense of explainability and contestability	Integrative perspectives and sense-making across both concepts	46
Perspectives Between Fields	Cross-disciplinary views and comparative insights	146
Perspectives Between Fields: Complexity	How complexity differs across disciplines and impacts analysis	21
Perspectives Between Fields: Conceptual blending	Integrating concepts from multiple fields	6
Perspectives Between Fields: Connecting fields	Bridging disciplinary approaches and knowledge	43
Perspectives Between Fields: Decoupling	Separating concepts or practices across fields	6
Perspectives Between Fields: Disparities between fields	Differences in methods, goals, or assumptions	26
Perspectives Between Fields: Human judgment versus machine judgment	Comparative evaluation of human and machine decision-making	9
Perspectives Between Fields: No use	Claims that certain cross-field ideas are not useful	9
Perspectives Between Fields: Normative versus technical	Tensions between value-driven and technical perspectives	10
Perspectives Between Fields: Polysemy	Multiple meanings of key terms across fields	16
Regulation	Legal frameworks, policies, and governance of AI systems	219
Regulation: challenges	Practical and conceptual hurdles in regulation	26
Regulation: citizen's rights	Rights of individuals affected by AI (e.g., appeal, access)	15
Regulation: definition	Attempts to define regulatory scope and terms	7
Regulation: elements	Key components and instruments of AI regulation	24
Regulation: fairness	Regulatory approaches to fairness and bias	19
Regulation: Flawed policy	Critiques of misguided or poorly designed regulations	15
Regulation: Global politics	International dynamics influencing AI regulation	2
Regulation: implementation	How regulations are applied and enforced	57
Regulation: Judicial flexibility	Flexibility and discretion within judicial processes	3
Regulation: Principle- vs. rule-based	Comparing high-level principles to specific rules	3
Regulation: proportionality	Balancing measures relative to risks and impacts	5
Regulation: purposes	Goals and rationales for regulating AI	14
Regulation: Spirit of the Law	Interpreting and applying underlying legal intent	7
Regulation: Standardization process	Role of standards bodies and technical norms	3
Regulation: suggestions	Proposed regulatory improvements and reforms	8
Regulation: Unclarities	Ambiguities and uncertainties in regulatory text or practice	11
Stakeholder considerations	Views, needs, and roles of affected stakeholders	141
Stakeholder considerations: Adapting to the algorithm	How stakeholders adjust behavior to algorithmic systems	4
Stakeholder considerations: Citizens' needs	Public needs, preferences, and expectations	49
Stakeholder considerations: empower stakeholders	Strategies to empower or give voice to stakeholders	4
Stakeholder considerations: human in the loop	Roles and responsibilities of humans in decision chains	4
Stakeholder considerations: Overburdened Human in the Loop	Risks of excessive workload or responsibility	7
Stakeholder considerations: Participation challenges	Barriers to meaningful stakeholder participation	5
Stakeholder considerations: stakeholder needs	Specific needs identified across stakeholder groups	4
Stakeholder considerations: Stakeholder responsibilities	Duties and obligations across stakeholder roles	18
Stakeholder considerations: Stakeholder roles	Typologies and functions of different stakeholders	37
Stakeholder considerations: Who is the user?	Clarifying user definitions and primary beneficiaries	8
Technical implementation	Practical technical methods, tooling, and deployment practices	24
Value sensitive design	Designing with human values integrated throughout	11