

“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, Department of Sustainable Design Engineering, The Netherlands

KARS ALFRINK, Delft University of Technology, Department of Sustainable Design Engineering, The Netherlands

THOMAS LE GOFF, i3, Télécom Paris, Institut Polytechnique de Paris, France

TIPHAINE VIARD, i3, Télécom Paris, Institut Polytechnique de Paris, France

Explainability and its emerging counterpart contestability have become important normative and design principles for the trustworthy use of AI as they enable users and subjects to understand and challenge AI decisions. However, the regulation of AI systems spans technical, legal, and organizational dimensions, producing a multiplicity in meaning that complicates the implementation of explainability and contestability due to the difficulty of defining them. Resolving this conceptual ambiguity requires specifying and comparing the meaning of both principles across regulation dimensions, disciplines, and actors. This process, here defined as *translation*, is essential to provide guidance on the principles’ realization. To this end, we present the findings of a semi-structured interview study with 14 interdisciplinary AI regulation experts. We report on the experts’ understanding of the intersection between explainability and contestability in public AI regulation, their advice for a decision subject and a public agency in a welfare allocation AI use case, and their perspectives on the connections and gaps within the research landscape. We provide differentiations between descriptive and normative explainability, judicial and non-judicial channels of contestation, and individual and collective contestation action. We further outline three main translation processes pertaining to the alignment of top-down and bottom-up regulation, the assignment of responsibility for interpreting regulations, and the establishment of interdisciplinary collaboration. Our contributions include an empirically grounded conceptualization of the intersection between explainability and contestability and recommendations on implementing these principles in public institutions. We believe our contributions can inform policy-making and regulation of these core principles and enable more effective and equitable design, development, and deployment of trustworthy public AI systems.

Additional Key Words and Phrases: explainability, contestability, regulation, AI, interdisciplinary research, qualitative methods

1 Introduction

Explainability and contestability are central principles in the trustworthy development and deployment of public AI systems [104]. However, while these principles are defined and discussed in both explainable AI (XAI) [47, 59, 102] and legal [7, 52, 65] research, an approach connecting these perspectives has not yet been adopted. The need for this 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 contestability¹, they do not include guidelines to translate their legal provisions into tangible system requirements [65].

¹The right to explanation is debated [22, 89, 98], but texts such as the GDPR, DSA, and EU AI Act provide for explainability in algorithmic decisions [65].

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@tudelft.nl, Delft University of Technology, Department of Sustainable Design Engineering, Delft, The Netherlands; **Kars Alfrink**, c.p.alfrink@tudelft.nl, Delft University of Technology, Department of Sustainable Design Engineering, Delft, The Netherlands; **Thomas Le Goff**, thomas.legoff@telecom-paris.fr, i3, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France; **Tiphaine Viard**, tiphaine.viard@telecom-paris.fr, i3, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France.

Explainability serves the purpose of increasing stakeholders’ understanding of an AI system and to enable informed decision-making [55], while contestability allows stakeholders to challenge and appeal algorithmic decisions [62]. A range of methods (‘mechanisms’ [4]) has been suggested in XAI and HCI through which these principles are realized, such as explaining feature importance, providing counterfactuals, or requesting intervention. 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.

We formulate two challenges that impede the implementation of explainability and contestability. First, both explainability and contestability are polysemic 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 [48], domain [99], and use-case setting [15]. This multiplicity of meaning excludes a one-size-fits-all approach [41, 75] and instead requires guidelines that can be applied in a variety of contexts [65]. And second, AI in high-stakes scenarios is a comparably new phenomenon, with both theoretical background and regulatory oversight still in development. Consequently, there are few best practices and guidance that can aid in the implementation of contestability [5, 52, 62] and interdisciplinary approaches to the creation of legislation have only begun to be mapped out [70]. Closing this gap between regulation and implementation requires policy-making that is evidence-informed [63], i.e., that is supported by research that bridges disciplines and provides empirical grounding.

To this end, we present findings from a task-based interview study with 14 interdisciplinary experts in AI regulation. Participants engaged in a card-sorting activity, analyzed a welfare allocation AI use case, and reflected on a graph representing the research landscape of explainability and contestability. Our findings highlight distinctions between descriptive and normative explainability, judicial and non-judicial contestation channels, and individual and collective contestation action. Participants further defined the intersection of explainability and contestability in their capacity to improve citizen empowerment and highlighted that both principles are not effective if the underlying policy governing a system’s deployment is flawed. In the analysis of our findings, we are guided by the following research questions.

[RQ1] *Conception*: How do AI regulation experts define and map the intersection between explainability and contestability of AI systems?

[RQ2] *Implementation*: Which are the major points of translation and points of friction in the (regulatory, institutional, and technical) implementation of explainability and contestability?

This work contributes an empirically grounded differentiation of explainability and contestability, a description of the major processes of translations and points of friction in their implementation, and a discussion embedding both principles into the larger research landscape. Our objective is to support an evidence-informed [21, 63] discussion to guide the implementation of policies surrounding the deployment of explainable and contestable public AI systems.

2 Background and related work

In this section, we situate our work in relation to public sector AI, introduce the principles of explainability and contestability, and describe the work to date on their intersection, both from a policy and a design perspective. We then conceptualize regulation and translation, two lenses that we will use to guide our analysis.

2.1 Trustworthy AI systems in the public sector

This study examines high-risk AI systems deployed in public institutions that can significantly impact individuals’ fundamental rights, safety, or health [50]. Research shows that these systems, despite their deployment in high-risk domains, are frequently dysfunctional [54, 79], discriminatory [24], and harmful through aggravating power imbalance [34, 72] and restricting autonomy [78]. For these reasons, both researchers [9, 13, 49, 84] and policy makers [35, 46, 74, 75] have advocated that high-risk AI systems should align with value frameworks such as *trustworthy* or *responsible* AI [27, 40, 94], which emphasize human agency, oversight, transparency, accountability, and fairness [39]. Explainability and contestability support these frameworks by enabling people to understand [55] and challenge [44] AI decisions. Although both principles are integrated into various design frameworks [5, 48, 55, 59], their implementation as part of the EU AI regulation remains challenging [38, 65, 70], as we describe in Section 2.2.3.

2.2 Explainability and contestability in policy and design

2.2.1 Explainability. Policy texts understand explainability to provide information about an AI system’s logic [36], core parameters [37], and specific purposes [8], while allowing users to comprehend system operations [65] and interpret results [35]. Although established as a principle, these texts leave implementation details unspecified, including choices between global or local explanations and whether they should be provided before or after decisions [65]. These details can make large differences in the design of explanations since the aim for developers [57, 58, 67, 81] is fundamentally different from that of non-technical stakeholders [6, 48, 66]. Effective explanations adapt to stakeholders’ expertise [18, 66] and objectives [41, 48], whether it is evaluating fairness [32, 91] or understanding the deployment context [33, 88].

2.2.2 Contestability. Contestation rights are contained in the EU Charter of Fundamental Rights’ guarantee of being heard [14] and in both GDPR Article 22 and the Council of Europe’s Convention 108+² [7, 30]. Research on contestable AI design examines how various stakeholders, from human operators to decision subjects and third parties, can utilize mechanisms to challenge algorithmic decisions [4, 7, 60]. Examples of such mechanisms include data input control [56], decision revision requests [59], and various audit methodologies [29, 64, 80, 90]. In both design [3, 7] and policy frameworks [22, 52, 65], the principle of contestability is described to be enabled through explainability.

2.2.3 Connecting explainability and contestability. Explanations are considered foundational for effective contestation, regarding both individual decisions [96] and system-level performance [23]. Understanding the intersection of both principles is thus crucial for the development of trustworthy public AI systems that support human autonomy, justice, and legitimacy [28, 44, 56]. However, there is no established design approach to realize the rights to contestation and explanation in AI systems [52, 89]. Importantly, given that current regulations lack clear implementation guidelines and that judicial systems cannot accommodate universal contestation rights [52], both explainability and contestability will manifest themselves largely through the actions of regulated actors and extrajudicial processes. This means that their operationalization through mechanisms [70] and human oversight configurations [31] will be determined by public institutions, standardization bodies [65], industry [28], and civil society organizations [71]. This interpretative flexibility, combined with competing interests, leads to diverging perspectives on how both principles should be implemented according to regulation.

Empirical research at the intersection of explainability and contestability has revealed nuanced relationships between these principles. Previous research has shown that explanations supported participants’ perceived *informational*

²The Council of Europe’s Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data.

fairness and contestation mechanisms enhanced *procedural* fairness, but that explanations alone did not prompt participants to question the decision-making process overall [102]. In contrast, another study demonstrated that counterfactual explanations indeed led participants to challenge the underlying decision system by revealing structural inconsistencies [96]. Lastly, in design workshops on contestation mechanisms, participants proposed features that often aligned more closely with explainability than with contestability [5]. While the delineation between explainability and contestability thus remains challenging, research consistently indicates that stakeholders value mechanisms that allow active influence, incorporating principles of autonomy [78], empowerment [28], and representation [95].

2.3 Analytical lenses: Regulation and Translation

In the following, we propose two lenses to explore the relationship between explainability and contestability: regulation and translation. The regulatory lens reveals how formal requirements and design choices shape these principles, while the translation lens examines how diverse fields—such as computer science, policy-making, and design—interpret and operationalize them. By analyzing stakeholder disagreements and key decision-making bottlenecks, we aim to shed light on how these principles are realized in practice.

2.3.1 The role of regulation in achieving trustworthy public AI. Black [11] characterizes regulation as “a process involving the sustained and focused attempt to alter the behavior of others according to defined standards or purposes with the intention of producing a broadly defined outcome”. This process becomes particularly complex when regulating immaterial sociotechnical artifacts such as AI systems, rather than traditional material objects. This is due to the multidisciplinary nature of AI, which poses unique challenges for regulatory frameworks in balancing technological capabilities, societal values, and ethical principles.

Although classical regulation relied on unilateral “command-and-control” mechanisms by public institutions [17], modern approaches have evolved towards “regulation by design” [68, 101], where stakeholders collectively translate legal norms into technology and actionable practices, creating a feedback loop between regulatory intent and implementation [77]. Importantly, internal governance and design choices within public institutions thus constitute their own regulatory dimension. Public AI systems, for example, embed societal values through design decisions that shape citizen interaction [69]. Effective AI regulation must, therefore, reconcile top-down normative frameworks with bottom-up processes driven by technological design and user interactions, necessitating an understanding of the interdependence of legal frameworks, design practices, and institutional governance.

2.3.2 Examining explainability and contestability through their processes of translation. The sociology of translation studies the ways in which diverse groups of actors and organizations make sense of concepts such as explainability and contestability in the act of regulating public AI. Social worlds, first theorized by Becker [10], encompass the actors and organizations that shape a given concept. Understanding these social worlds is crucial for analyzing how individuals and organizations collaborate and compete and how knowledge production is structured. Science and technology studies (STS) have examined technical objects in society through frameworks such as the Social Construction of Technology [76] and Actor-Network Theory [19]. These approaches posit that technological objects and their societal deployment are inseparable, each mediating the other’s development and impact.

2.3.3 Using regulation and translation to understand mobile phones—an example. A typical example of a sociotechnical system is the mobile phone: while it is a concrete technological object, the mobile phone is mostly useless without regulatory frameworks governing frequency allocation, infrastructure deployment, and data privacy. These regulations

shape both technical standards and user behavior. Simultaneously, different social worlds translate the mobile phone’s meaning and purpose differently: telecommunications engineers see it as a network node requiring optimization, privacy advocates as a surveillance risk, teenagers as a social status symbol, and businesses as a productivity tool. This multiplicity of interpretations influences how various stakeholders engage with mobile phone regulation, from spectrum allocation to privacy protection measures.

3 Methods

In this section, we describe our methods and study procedure. We conducted a task-based interview study with 14 experts in the regulation and design of AI systems (Section 3.2). The interviews were composed of three main elements: a card sorting activity, a use case discussion, and an analysis of a citation graph representing research on explainable and contestable AI (Section 3.1). The interviews were analyzed using inductive and deductive thematic analysis (Section 3.3).

3.1 Study setup and procedure

Figure 1 gives an overview of the overall study process. All interviews were conducted online using the collaboration platform *Miro* and took around 90 minutes.

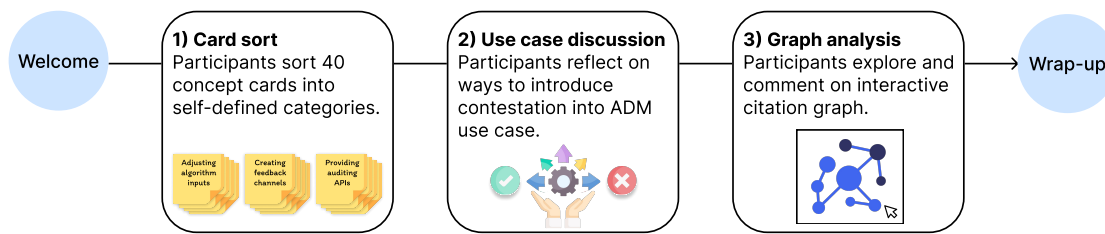


Fig. 1. **Study procedure overview.** Participants completed three sequential tasks: 1) sorting 40 concept cards into self-defined categories related to AI systems, 2) analyzing explainability and contestability in a specific algorithmic use case from personal and institutional perspectives, and 3) exploring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

3.1.1 Card sort. In the first activity, participants were asked to sort 40 cards containing explanation and contestation mechanisms into self-defined categories to elicit their mental structure and understanding of both principles. Card sorts are an empirical method of uncovering how people organize and categorize knowledge [100], in which participants sort a set of cards into categories that can be defined a priori (closed) or by the participants themselves (open) [86]. When the sorting is complete, each category is named and discussed. The method allows comparing differences and commonalities in how people perceive and categorize the items in question [85]. The recommended number of cards depends on the variety of card types and the types of items and ranges from 20 to 100 cards [85, 93].

For the study, we conducted an open card sort with 40 cards where participants sorted the cards only once (“all-in-one” approach [85]). The open sort allowed categorization to emerge organically, avoiding predefined categories. The single sort was chosen due to insights from four pilot studies, which indicated that repeated card sorts would prolong the study duration without giving many new insights. The set of cards represented a range of explanation and contestation mechanisms derived from publications in human-computer interaction, legal studies, and design research [5, 25, 56, 59, 61, 62, 65, 95, 98, 102]. These papers were selected on the basis of citation counts and reference

networks and were used to extract processes and design solutions to create the cards. Participants could add items using spare cards if needed. Each card featured a gerund phrase (e.g., ‘Evaluating individual results,’ ‘Disclosing the algorithm to experts,’ ‘Requesting new decisions’) to emphasize action while remaining flexible for interpretation. A depiction of the full set of cards is provided in the supplementary material.

3.1.2 Use case discussion. To examine how participants would apply their understanding of explainability and contestability in practice, the second part of the study asked participants to review a description of an algorithmic system designed to predict welfare fraud, detailing its purpose, training data, and public reception. They were presented with two scenarios: a fictional welfare beneficiary who was flagged for fraud and wanted to contest the decision, and the deploying social security agency looking for advice to enhance the system’s explainability and contestability. Participants were asked to provide advice based on existing measures in similar contexts and their professional experience. These scenarios aimed to encourage participants to consider different perspectives and address all relevant aspects of the case.

The study’s use case represented a public AI system used by *Caisse d’allocations familiale* (CAF), a part of the French social security services [83]. The model uses logistic regression to predict welfare fraud likelihood among beneficiaries and was trained on data from household investigations and corresponding overpayments. The number of decision subjects amounts to 13 million households (nearly half of France’s population), of which 100,000 are annually flagged for detailed investigation [82]. 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.

3.1.3 Citation graph discussion. To elicit participants’ thoughts on the landscape and interconnection of academic literature regarding explainability and contestability, we asked them to comment on a network representation of said literature. 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. We included the latter to create links with explainability in legal literature while excluding unrelated work in XAI. This query returned 648 papers, spanning from 1991 to 2025; note, however, that 151 papers (23.3%) have been published in 2024 alone, and 465 (71.8%) have been published in 2020 or after. We built the network $G = (V, E)$ from these results; the set of nodes V represented the articles, an edge was inserted between two articles if they shared at least $k = 4$ references, i.e., they cited the same bodies of work. We set k manually after iterating through values between 1 and 10, selecting the value that increased visual readability.³ In order to exhibit clusters in the graph, we used the Louvain algorithm [12]. The graph was presented interactively, allowing the interviewees to get more details (e.g., title, year of publication, authors, ...). The graph is available under contest.graphuzu.fr and depicted in the appendix.

³At the extremes, $k = 10$ leads to multiple disconnected subgraphs, and $k = 1$ leads to an overly dense graph that is hard to read visually.

3.2 Participants

We recruited 14 experts from European universities and public institutions who could provide insight into the study’s research questions through their research or occupation. Participants were selected based on their professional background, examples of which included involvement in the EU Commission’s High-Level Expert Group on AI (EU AI HLEG), in standardization bodies’ regulation processes, or the legislative process of the EU AI Act and similar legal texts.⁴ This selection of participants aimed to capture interdisciplinary viewpoints on explainability and contestability, the main aspects of their regulation, and the steps and challenges public institutions would face in their implementation. The recruitment methods included recommendations from the authors’ networks, snowball sampling, and direct invitations. The sample size followed qualitative research guidelines, focusing on code and meaning saturation [45].

3.3 Analysis

For RQ1-Conception, we used inductive thematic analysis [16] to examine the participants’ conception of explainability and contestability, the main aspects of their regulation and institutional implementation, the role and responsibilities of the involved stakeholders, and the differences in perspectives between disciplines. The created codes were regularly compared between authors to develop, merge, and delete codes. RQ1-Conception is further informed by comparisons of how participants arranged and categorized explanation and contestation mechanisms in the card sorts, focusing particularly on the sorting criteria and thematical similarities between participants’ clusters.

For RQ2-Implementation, we employed both inductive and deductive thematic analysis to identify the main processes of translation and points of friction in participants’ reports. Prifti et al. [77]’s categorization of regulation by design here served as a framework to structure the main elements and processes of regulation, as it covers both top-down and bottom-up approaches as described in Section 2.3.1. Further, to capture the notion of translation and map it to the mentioned regulation structure, we drew from Callon [20]’s work on translation in sociology.

4 Results

In this section, we describe our results on how AI regulation experts conceptualize and understand the principles of explainable and contestable AI (RQ1), as well as the major processes of translations and points of friction in the regulatory, institutional, and technical implementation of these principles (RQ2).

4.1 RQ1-Conception: How do AI regulation experts define and map the intersection between explainability and contestability of AI systems?

4.1.1 Explainability and contestability are polysemic and thus require differentiation. We find that while participants had similar understandings of the overall principles of explainability and contestability, their conceptions differed in how these principles can be differentiated and applied in practice. Overall, **explainability** was 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). Adapting formulations from the interviews, we define the first kind as **descriptive explainability** and the second kind as **normative explainability**. Participants further linked this distinction to the definitional differences between disciplines, as illustrated by the nature of *justifications* and *reason-giving*:

⁴To allow participants the freedom of anonymous expression in the interviews, demographic as well as occupational details on their persons were omitted.

A core difference between the two sides is in terms of what you expect an explanation to do. [...] [L]egal literature expects explanations to be something that's more a reason-giving on the decision process [normative], but then you have the problem of understanding how far the current techniques take us [descriptive]. (P9)

Contestability was primarily conceived as allowing stakeholders to challenge AI decisions. Examples included enabling regulators to “critically process the information provided to them, and also push back against it” (P6) and affected persons to “understand the situation and to file complaints” (P6). Participants further introduced two key distinctions in how contestation is realized: The first distinction is between **collective action** and **individual action**. Whereas individual contestation affords subjects the means to challenge AI decisions that affect them personally, collective contestation means “translating personal issues into general matters and public fights” (P12) by “contesting decision patterns rather than individual decisions” (P9). The second distinction is between **judicial channels** and **non-judicial channels**. 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 through a mediation system or ombudspersons. Participants emphasized the importance of differentiating the available channels of contestation and the type of action 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. [I]ndividual contestation [is seen] as prejudicial, [...] something that can be useful at times, but this is the main concern. (P9)

4.1.2 *The intersection of explainability and contestability can be mapped through their goals, mechanisms, and limits.* The question of how both principles are linked was a central topic in the interviews with AI regulation experts. Participants often posited explainability to be a prerequisite to contestability, stating “contesting presupposes understanding” (P2), and, more directly, “it’s not that we want explainability for its own sake, but [...] because it facilitates contestability” (P4). In the following, we report on how and why these two principles were perceived as tightly linked.

Goals. Participants frequently described explainability and contestability through the fact that they worked towards the same goals and purposes. These goals were summarized as allowing “full human agency” (P2), supporting “the rule of law” (P4), and enabling people to “gain more control in a decision that is important to them” (P11). They were also understood to **benefit citizen empowerment** by alleviating opacity and information asymmetry:

The problem is that we citizens, we don’t know the technologies used by the public administrations. And therefore, we are in a blind world in which we cannot contest, because we don’t know what is used and to what purpose and how it works. (P3)

Explanations serve the goal of citizen empowerment by enabling citizens to assess two distinct aspects: the **acceptability of individual decisions** and the **overall suitability of a system’s deployment**. This dual assessment can, in turn, prompt actions of contestation at the individual or collective level. In the study’s use case, participants explained that an AI system might not be contestable in its individual decisions if all rules are followed; but that global explanations of decision patterns can enable contesting the entire AI system if, e.g., the underlying policy is identified as dysfunctional. In contrast, the acceptability of individual decisions becomes crucial when subjects face unfavorable outcomes and must decide whether to contest them. P9 noted that acceptability is important so that citizens would not use contestation channels only to impede the process and that citizens have a “good faith obligation” to “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).

For these reasons, **preventing the obstruction and gaming of decision processes** was perceived as a side goal when implementing explanation and contestation measures. In their advice for a social security agency on implementing both principles, participants stated concern that: *“If they do it properly, they will be submerged by requests and contestation”* (P12). This was explained in three main points: i) citizens want to take control over the decision process to achieve favorable results, ii) if citizens feel that their values are not represented in the decision-making process, they will be more likely to contest, and iii) while citizens should not be overburdened with responsibilities, fully withdrawing them and thus all ways of control forces citizens to *“adapt their behavior to their interpretation of the algorithm”* (P5). This is depicted in Figure 2-A, in which spheres of responsibility of civil servants and developers are clearly separated from those of the decision subject. As a remedy to the imbalance of responsibilities, participants suggested to **include citizens in the design process** of public AI systems to improve acceptability and further fulfill requirements obliging administrations *“to consult the population in terms of impact and gather public feedback”* (P9).

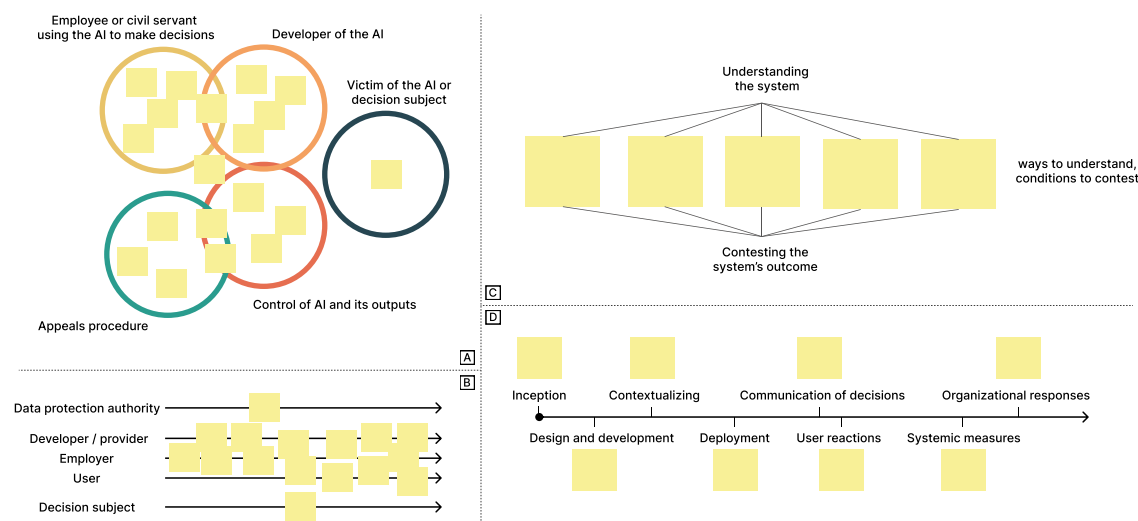


Fig. 2. **Card sort structures.** The image shows the structure of four completed card sorts. Participants chose different criteria and dimensions to sort the cards into clusters, including spheres of responsibility (A), responsibility over time and per role (B), ways in which mechanisms connect both to understanding and contesting (C), and an allocation to the implementation process over time (D).

Mechanisms. Although all mechanisms discussed in the interviews could be argued to relate to both explainability and contestability, participants showed a more detailed understanding of mechanisms that supported explainability than contestability. During the card sort activity, participants often created a cluster of cards that was titled with ‘transparency’, ‘understanding the system’, ‘explanations’, or ‘foundations of what you’re dealing with’. A selection of examples is depicted in Figure 3.

In contrast, participants rarely created clusters that were connected to the overarching principle of contestability, but rather clusters that covered different aspects of it, such as ‘control’, ‘appeals procedure’, ‘rectification’, ‘judicial remedies’, and ‘auditing’. These categories partly mirror the distinctions described in Section 4.1.1 (judicial vs. non-judicial channels, individual vs. collective action), indicating that contestation was not perceived as a homogenous and self-contained collection of mechanisms and more as a principle that is realized differently depending on the contestation

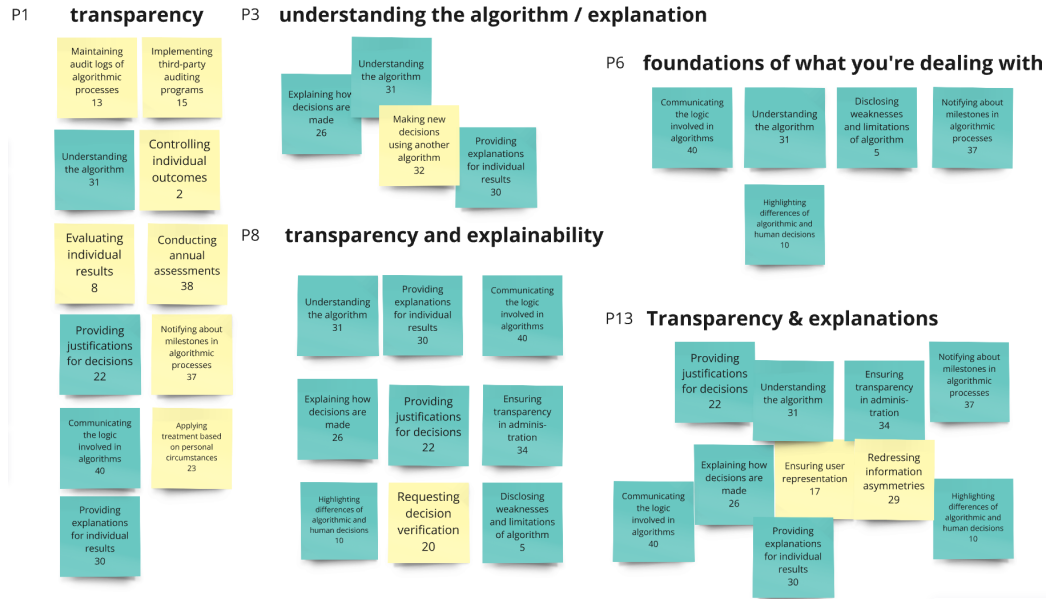


Fig. 3. **Card sort explainability clusters.** The image shows parts of six different card sorts, all containing cards that participants related to transparency or explainability. Often, participants created exactly one cluster with a similar name to those shown. Cards that can be found in multiple of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

actor, subject, goals, and channels. Exceptionally, P2’s card sort contained both a specific contestability cluster and designated a number of mechanisms as “ways to understand, conditions to contest” (P2), represented in Figure 2-C.

We thus see the cause for the diverse interpretation of contestation mechanisms in the multiplicity of ways to realize them, but also in the fact that participants were not always well-acquainted with design mechanisms that support contestability (described in Section 4.1.3). P6 further explained that the mechanisms relating to transparency and explanation could be comparably more well-known because the regulation of data protection is so prevalent:

[D]ata protection law is such a heavy hammer that is being wielded very frequently because [...] everybody has heard about [it]. And that means that sometimes people try to address issues with data protection law instead of going other routes, which maybe would make more sense if you stop to think about that. (P6)

Limits. Participants repeatedly described circumstances in which explainability and contestability would not be enough to resolve the core issues of information asymmetry and power imbalance. Three key considerations were that explanation and contestation mechanisms are ineffective if i) the system’s actual purpose is to enforce sanctions on decision subjects, ii) the sampling strategy targets vulnerable populations disproportionately, and iii) non-judicial channels of contestation are obscured or not available, as then “you cannot even exercise [the right to contestation] properly without assistance” (P9). P4 elaborated on these aspects when asked which advice they would give to a decision subject in the face of an unfavorable outcome:

[...] [S]he doesn’t stand a chance. Because the whole system is biased against people like her due to the indicators that are selected for the model. Secondly, she can hardly do anything about those indicators herself.

[...] The actual problem [...] [is] not only that the system is simply badly designed, it was badly designed on purpose. (P4)

Participants further explained that these design choices do not necessarily originate from the agency itself but can be prescribed “*from the political level*” (P5), thus confining the agency to the implementation of values that are not self-chosen nor aligned with those of decision subjects. Several participants emphasized that individual explanation and contestation action would be ineffective before systemic issues regarding values and policy decisions were resolved.

Regarding individual contestation, participants further problematized that it **primarily serves individual interests** and thus has less potential to effect change on a systemic level. Based on their experience, P12 elaborated that decision subjects “*want to fight for **their** privacy and **their** freedom. It’s very different from fighting for privacy with a big P in general*” (P12). For these reasons, collective contestation was posited as a more effective alternative for “*contesting decision patterns [...] through things like class actions*” (P9). Enacting this contestation using judicial (e.g., class action lawsuits) and non-judicial channels (e.g., civil society organizations), which facilitate “*translating personal issues into general matters and public fights and political debates that we should have*” (P12), was noted as being insufficiently explored within the current EU jurisdiction. The **capacity of inciting political debates** was further assigned to contestability rather than explainability for two main reasons: First, because for “*neural nets or very complex decision trees [...] you can’t come up with an explanation at all*” (P8); and second, because a descriptive explanation does not justify that “*the decision is correct, accurate, and legit; you need to justify so people are convinced by it*” (P11), thus lacking the normative force that is essential for contestability. Leveraging the distinction described in Section 4.1.1, justification might thus be defined as a limit of *descriptive* explainability, but as an essential part of *normative* explainability.

4.1.3 Gaps between disciplines impede mapping the intersection of explainability and contestability. Participants were often well-acquainted with the works on the “right to explanation” [22, 89, 98] and with judicial ways of contestation, such as appeals and redress, but comparably unfamiliar with the term “contestability” and the corresponding body of work in design and HCI research. Several participants commented on this, stating “*I’m thinking in which world I was living [...], I didn’t notice that it 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: “*I think that this should be much more an explosion of different colors. [...] 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 definitory and conceptual separation between the disciplines is one of the main challenges in mapping the intersection between explainability and contestability.

4.2 RQ2-Implementation: Which are the major points of translation and points of friction in the (regulatory, institutional, and technical) implementation of explainability and contestability?

4.2.1 Practical implementations of explainability and contestability should keep in mind the spirit of the law.

Translation. 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 EU AI HLEG and took part in the conception of the EU AI Act described that the group was tasked with identifying the ethical principles on which the ethics guidelines should be based. Taking inspiration from biomedical

ethics, these principles included autonomy, justice, non-maleficence, and explainability, which replaced beneficence, which are common topics in ethical AI charters [51]. Other participants who were also part of the EU AI HLEG confirmed this consensus on explainability, stating that “*lawyers agreed in there, and human rights experts agreed in there, practitioners, so that was a consensus among 55 or 52 people, and there was no doubt that this is a fundamental requirement*” (P8). Importantly, explainability was 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 really important to have this because it also [...] reflects on [...] contestability*” (P10). However, this fundamental requirement, the spirit of the AI Act with regard to explainability, is subjected to a number of transformations before arriving at practical applications: its formulation as legal texts and as technical standards, its integration into national jurisdiction, and, finally, its implementation in public institutions.

Point of friction. Participants highlighted two main concerns that could interfere with this process. First, they 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 whenever [...] a software application could be a high-risk application. [...] [W]hen I was listening to them, I thought, no, this was not the intention of the entire endeavor here. It was just [...] to safeguard certain principles*” (P8). This perception was shared by other participants, who 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). Second, public perceptions of the EU AI Act created by the media were identified as a major obstacle in its implementation, described as the “*hype aspect [...], because it takes so much time and energy to just get back down to reality and make sure that everybody has the same grounding for these conversations*” (P6).

4.2.2 The implementation of explainability and contestability explicates how responsibility moves between regulators.

Translation. Participants repeatedly reflected on who would be responsible for implementing explanation and contestation mechanisms (as depicted in Figure 2-A and B). Participants considered ‘regulators’ [77], including policymakers, standardization bodies, data protection authorities, and developers, to be the main actors in this translation operation. The translation itself is driven through the “*shared responsibility to ensure user representation in the development and the use of the AI*” (P5) and the fact that “*what the people are struggling with is at very different levels [...], I mean, what does explainable AI mean?*” (P8). Technical standards were perceived to be one of the main components to clarify the allocation of responsibility, but their enforcement again raised questions:

[I]f there are decisions made by a software company, upon services or somehow affecting [people’s] rights, this needs to be justified and explicable. And if you can’t explain why a certain decision was made or a service was not offered, you have a problem. (P8)

Point of friction. Due to the conceptual ambiguity 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, to my knowledge. 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 translation operation 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).

4.2.3 Collaboration between communities is painful but strengthens the implementation of explainability and contestability.

Translation. Participants who had come into contact with both legal and design research on AI regulation regularly highlighted the potential and shortcomings of interdisciplinary collaboration. To exemplify, 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 translation between the social worlds of disciplines, since “*as soon as we start to explain what we are doing and what we find out in our research to lay people, everyone else, also from different disciplines, could understand*” (P8). In turn, 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)

Point of friction. While the advantages of this translation process thus become evident, following through on it was described to be “*painful*” (P8). Especially for people removed from the technical sides of AI, “*even explaining the concept of explainability sometimes can be challenging because they have to understand that AI is a black box*” (P14), which is complicated further in the case of complex models that are difficult to explain in general. Similar comments were made about the distinction between judicial and non-judicial contestation actions, since “*there wasn’t really a distinction [...] [s]o that would have been like one big grouping, looking at it from a legislator’s perspective*” (P5). Participants who had experience in both technical and legal disciplines described it as “*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). Further, “*interdisciplinary research requires [...], a different type of language [and] [t]here is incommensurability between methods and approaches, [...] but I think that this is how we can resolve our grand challenges*” (P8). Finding translation operations that facilitate shared conceptual understanding and vocabulary is thus essential to effectively inform regulatory efforts.

5 Discussion

Drawing from the insights generated in the interviews, we suggest three main recommendations for policy-making and regulation in explainability and contestability. A figure visualizing the duality of both principles is included in the supplementary material.

Strengthening the intersection of explainability and contestability in legal instruments and trustworthy AI governance. In the interviews, a consensus emerged that contestability allows individuals to exercise control over AI usage by public institutions and that is based on information embodied by explainability. While individual action was seen as a way to assess the acceptability of specific decisions, collective action was considered more suitable for challenging the system’s overall suitability. Legal research suggests that the optimal governance scheme combines indirect control (by a regulation authority) and direct oversight (by decision subjects who appeal decisions and get redress, also called “democratic control” [13]) [43, 92, 97]. We find that this direct oversight by decision subjects through contestability mechanisms is both (i) considered positively by AI regulation experts and (ii) not very familiar to them when it comes to concrete ways to implement it, possibly due to its absence in regulatory instruments like the EU AI Act. The equilibrium between indirect and direct control over public AI thus should be reconsidered, e.g., by giving more place to direct control using non-judicial contestation means. This can be supported by providing explanations that disclose the purpose of an AI system’s development and deployment (normative) rather than merely describing its

workings (descriptive) [13]. Public institutions should thus ensure that provided explanations are relevant and aligned with the recipient’s goals and level of knowledge [26, 73] by considering which kind of contestation they enable [89]. Further, public institutions can engage with the regulation ecosystem (e.g., AI Office, standardization bodies) to receive support in the alignment between regulatory intent of explainability and contestability and their implementation [77].

Going beyond judicial contestability in lawmaking and public policy. The majority of the interviewees were familiar with judicial means of contestation and unfamiliar with non-judicial ones, which explains why this aspect of contestability remained overlooked in regulatory initiatives until now [52]. 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 [42, 87]. While policy needs to decide *what* can be contested, *who* can contest and *who* is accountable, and *which types of review* should be put in place [60], design and HCI research has proposed 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 [62] and can articulate their voice in the process [102]. Public institutions can thus benefit from engaging with design and HCI researchers regarding the implementation of contestability in AI system design.

Using participatory approaches to resolve implementation challenges of contestability. The design of AI systems involves encoding legislation into software [103], meaning that design decisions about input features, data types, and human-AI interactions become policy decisions that are no longer delegated to public deliberation but rather to third-party developers [69]. Because important decisions are thus made early in the design process, participants emphasized the need to involve stakeholders, particularly decision subjects, through “early-stage deliberations” [53]. By integrating contestability as a co-designed and technical feature of the system itself rather than as a legal standard to meet, public institutions could align values embedded in AI systems with those of citizens. This might improve the acceptability of AI decisions while avoiding excessive contestation during operation. To this end, we highlight two main aspects of participatory approaches: i) the deliberations’ level of abstraction and ii) the participation mechanisms. Regarding the level of abstraction, prior work suggests that, instead of focusing on technical design decisions, participatory approaches should center around the values and policies embedded in code [2]. These values and policies can, in turn, be selected through citizen-wide participation before being embedded in AI systems [13]. Citizen assemblies or advocacy groups that represent groups that are unable to participate fully (e.g., children) can be alternatives to direct participation mechanisms [13]. Mediation through civil society organizations or ombudspersons can further serve as additional channels for collective contestation.

5.1 Limitations

Like any research, this study had limitations. First, the single 30-minute card sort, despite being refined through four pilot studies, likely influenced the depth of participants’ exploration and classification of the cards. Second, the citation graph presented was a simplified subset of literature on explainability and contestability and did not capture the whole research landscape, but served as a prompt for participants to share their interdisciplinary experiences. Third, the focus on an EU context for recruitment and analysis of regulation and implementation procedures excluded comparisons with other jurisdictions, which we highlight as a fruitful direction for future research.

6 Conclusion

In this work, we conceptualized explainability and contestability and their translation for implementation by interviewing 14 interdisciplinary experts on AI regulation. We provided distinctions to facilitate these translations, including between normative and descriptive explanations, individual and collective contestation action, and judicial and non-judicial contestation channels. We further described three main processes of translation pertaining to the preservation of the regulation’s spirit, the responsibility for interpreting regulatory provisions, and the essential but difficult 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 this work, we aimed to inform research and policy efforts that leverage explainability and contestability in the development of trustworthy public AI systems.

AI Usage Statement

During the final preparation of this manuscript, we utilized three AI-assisted tools for copy-editing: Claude 3.5 Opus (released by Anthropic in 2024), Writefull, and Grammarly. These tools were used solely to improve clarity and readability without altering the paper’s intellectual content, methodology, or findings.

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. <https://doi.org/10.1145/3630106.3658962>
- [3] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. 2023. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI ’23). Association for Computing Machinery, New York, NY, USA, Article 8, 16 pages. <https://doi.org/10.1145/3544548.3580984>
- [4] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by Design: Towards a Framework. *Minds and Machines* 33, 4 (01 Dec 2023), 613–639. <https://doi.org/10.1007/s11023-022-09611-z>
- [5] Kars Alfrink, Ianus Keller, Mireia Yurrita Semperena, Denis Buligin, 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] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Diaz-Rodriguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (Nov. 2023), 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- [7] 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. <https://doi.org/10.1145/3322640.3326699>
- [8] 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. <https://doi.org/doi/10.2864/28842>
- [9] 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.
- [10] Howard S Becker. 2023. *Art worlds: updated and expanded*. Univ of California Press.
- [11] Julia Black. 2001. Decentering Regulation: Understanding the Role of Regulation and Self-Regulation in a ‘Post-Regulatory’ World. *Current Legal Problems* 54, 1 (12 2001), 103–146. <https://doi.org/10.1093/clp/54.1.103> arXiv:<https://academic.oup.com/clp/article-pdf/54/1/103/7524076/54-1-103.pdf>
- [12] 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. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- [13] 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. <https://doi.org/10.1145/3630106.3658541>

- [14] Marco Borraccetti. 2011. *Fair Trial, Due Process and Rights of Defence in the EU Legal Order*. 95–107. https://doi.org/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. <https://doi.org/10.1145/3490099.3511139>
- [16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- [17] Stephen G. Breyer. 2009. *Regulation and Its Reform*. Harvard University Press, Cambridge.
- [18] Ruth M.J. Byrne. 2023. Good Explanations in Explainable Artificial Intelligence (XAI): Evidence from Human Explanatory Reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. Macau, SAR China, 6536–6544. <https://doi.org/10.24963/ijcai.2023/733>
- [19] Michel Callon. 1984. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. *The sociological review* 32, 1_suppl (1984), 196–233.
- [20] Michel Callon. 1984. Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review* 32, 1_suppl (May 1984), 196–233. <https://doi.org/10.1111/j.1467-954X.1984.tb00113.x>
- [21] Fred Carden. 2009. *Knowledge to policy: making the most of development research*. SAGE, Los Angeles, Calif.
- [22] 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). <https://doi.org/10.15779/Z38M32N986> Publisher: btlj.
- [23] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. 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, 2254–2272. <https://doi.org/10.1145/3630106.3659037>
- [24] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. <https://doi.org/10.1089/big.2016.0047>
- [25] Danielle Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Washington Law Review* 89 (March 2014), 1–33.
- [26] 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.
- [27] Mark Coeckelbergh. 2020. *AI ethics*. The MIT Press.
- [28] Robert Patrick Collins, Johan Redström, and Marco Rozendaal. 2024. The right to contestation: Towards repairing our interactions with algorithmic decision systems. (2024). <https://doi.org/10.57698/V18I1.06> Publisher: International Journal of Design.
- [29] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1571–1583. <https://doi.org/10.1145/3531146.3533213>
- [30] Council of Europe. [n. d.]. *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>
- [31] Rebecca Crotoft, 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>
- [32] Luca Deck, Jakob Schaeffer, 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. <https://doi.org/10.1145/3630106.3658990>
- [33] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. <https://doi.org/10.1145/3411764.3445188>
- [34] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press, Inc.
- [35] European Commission. 2024. Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations. .
- [36] European Parliament and Council of the European Union. [n. d.]. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. <https://data.europa.eu/eli/reg/2016/679/oj>
- [37] 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>
- [38] 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. <https://doi.org/10.1145/3461702.3462470>
- [39] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (June 2019), 261–262. <https://doi.org/10.1038/s42256-019-0055-y>
- [40] 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. <https://doi.org/10.1007/s11023-018-9482-5>

- [41] 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. https://doi.org/10.1007/978-3-031-44064-9_3
- [42] Mélanie Gornet and Winston Maxwell. 2024. The European approach to regulating AI through technical standards. *Internet Policy Review* 13, 3 (July 2024). <https://doi.org/10.14763/2024.3.1784>
- [43] John Graham, Timothy Wynne Plumptre, and Bruce Amos. 2003. *Principles for good governance in the 21st century*. Vol. 15. Institute on governance Ottawa.
- [44] Clément Henin and Daniel Le Métayer. 2022. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* 37, 4 (Dec. 2022), 1397–1410. <https://doi.org/10.1007/s00146-021-01251-8>
- [45] 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. <https://doi.org/10.1177/1049732316665344>
- [46] 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>
- [47] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, Edinburgh United Kingdom, 95–99. <https://doi.org/10.1145/3064663.3064703>
- [48] 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. <https://doi.org/10.3389/fcomp.2023.1117848>
- [49] Saffron Huang, Divya Siddharth, 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. <https://doi.org/10.1145/3630106.3658979>
- [50] 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. <https://doi.org/10.1109/MC.2023.3235712> Conference Name: Computer.
- [51] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.
- [52] Margot E Kaminski and Jennifer M Urban. 2021. The right to contest AI. *Columbia Law Review* 121, 7 (2021), 1957–2048.
- [53] 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. <https://doi.org/10.1145/3613904.3642849>
- [54] 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. <https://doi.org/10.1145/3630106.3659007>
- [55] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesting, 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. <https://doi.org/10.1016/j.artint.2021.103473>
- [56] 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.
- [57] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI ’20). Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3313831.3376590>
- [58] 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.
- [59] 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. <https://doi.org/10.1145/3593013.3594041>
- [60] 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. <https://doi.org/10.1145/3449180>
- [61] 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].
- [62] 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. <https://doi.org/10.1145/3491102.3517606>
- [63] 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. <https://doi.org/10.2760/374191>
- [64] Amelie Marian. 2023. Algorithmic Transparency and Accountability through Crowdsourcing: A Study of the NYC School Admission Lottery. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) (FAccT ’23). Association for Computing Machinery, New York, NY, USA, 434–443. <https://doi.org/10.1145/3593013.3594009>

- [65] 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). <https://doi.org/10.2139/ssrn.4520754>
- [66] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- [67] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). <https://christophm.github.io/interpretable-ml-book>
- [68] Bronwen Morgan and Karen Yeung. 2007. *An Introduction to Law and Regulation: Text and Materials*. Cambridge University Press.
- [69] Deirdre K Mulligan and Kenneth A Bamberger. 2019. Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ* 34 (2019), 773.
- [70] 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. <https://doi.org/10.1145/3630106.3659028>
- [71] Luca Nannini, Agathe Balayn, and Adam Leon Smith. 2023. Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1198–1212. <https://doi.org/10.1145/3593013.3594074>
- [72] 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. <https://doi.org/10.1145/3630106.3658915>
- [73] 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.
- [74] OECD. 2019. Recommendation of the Council on Artificial Intelligence. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- [75] 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. <https://doi.org/10.6028/NIST.IR.8312>
- [76] Trevor Pinch. 2012. The social construction of technology: A review. *Technological change* (2012), 17–35.
- [77] Kostina Pifti, 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. <https://doi.org/10.1007/s11023-024-09675-z>
- [78] Carina Prunkl. 2022. Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence* 4, 2 (Feb. 2022), 99–101. <https://doi.org/10.1038/s42256-022-00449-9>
- [79] 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. <https://doi.org/10.1145/3531146.3533158>
- [80] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 33–44. <https://doi.org/10.1145/3351095.3372873>
- [81] 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. <https://doi.org/10.1145/2939672.2939778>
- [82] Manon Romain, Adrien Senecat, Soizic Pénicaut, 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/>.
- [83] 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.
- [84] 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 May 2019), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- [85] 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.
- [86] Gordon Rugg and Marian Petre. 2007. *A gentle guide to research methods*. Open University Press.
- [87] 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. <https://doi.org/10.1177/1023263X1302000404>
- [88] 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]
- [89] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (Nov. 2017), 233–242. <https://doi.org/10.1093/idpl/ix022>
- [90] 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. <https://doi.org/10.1145/3479577>

- [91] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Enhancing Fairness Perception – Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople’s Fairness Perceptions of Algorithmic Decisions. *International Journal of Human-Computer Interaction* (July 2022), 1–28. <https://doi.org/10.1080/10447318.2022.2095705>
- [92] Nathalie A Smuha. 2021. Beyond the individual: governing AI’s societal harm. *Internet Policy Review* 10, 3 (2021).
- [93] Donna Spencer and Todd Warfel. 2004. Card sorting: a definitive guide. *Boxes and arrows* 2, 2004 (2004), 1–23.
- [94] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy Artificial Intelligence. *Electronic Markets* 31, 2 (June 2021), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- [95] 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. <https://doi.org/10.1145/3476059>
- [96] Peter M. VanNostrand, Dennis M. Hofmann, Lei Ma, and Elke A. Rundensteiner. 2024. Actionable Recourse for Automated Decisions: Examining the Effects of Counterfactual Explanation Type and Presentation on Lay User Understanding. 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, 1682–1700. <https://doi.org/10.1145/3630106.3658997>
- [97] 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).
- [98] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal* (2017). <https://doi.org/10.2139/ssrn.3063289>
- [99] 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. <https://doi.org/10.1145/3397481.3450650>
- [100] Jed R Wood and Larry E Wood. 2008. Card sorting: current practices and beyond. *Journal of Usability Studies* 4, 1 (2008), 1–6.
- [101] 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. https://doi.org/10.1007/978-94-007-6970-0_32
- [102] 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. <https://doi.org/10.1145/3544548.3581161>
- [103] 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
- [104] 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. <https://doi.org/10.1007/s00146-022-01480-5>

Supplementary material

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

A Card sort material and citation graph



Fig. 4. **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.

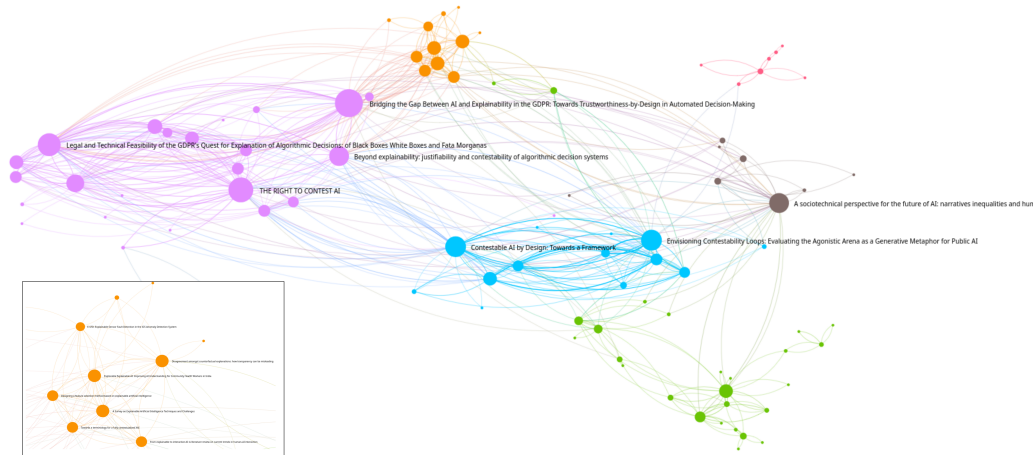


Fig. 5. **Overview of the citation graph.** The image shows the citation graph used in the study to elicit participants' reflections on the research landscape. The graph is the largest connected component of the co-reference graph related to contestability and AI, and it also includes references on the 'right to explanation'. The detail shows a zoom on the “explainable AI” cluster in the network.

B Duality of explainability and contestability

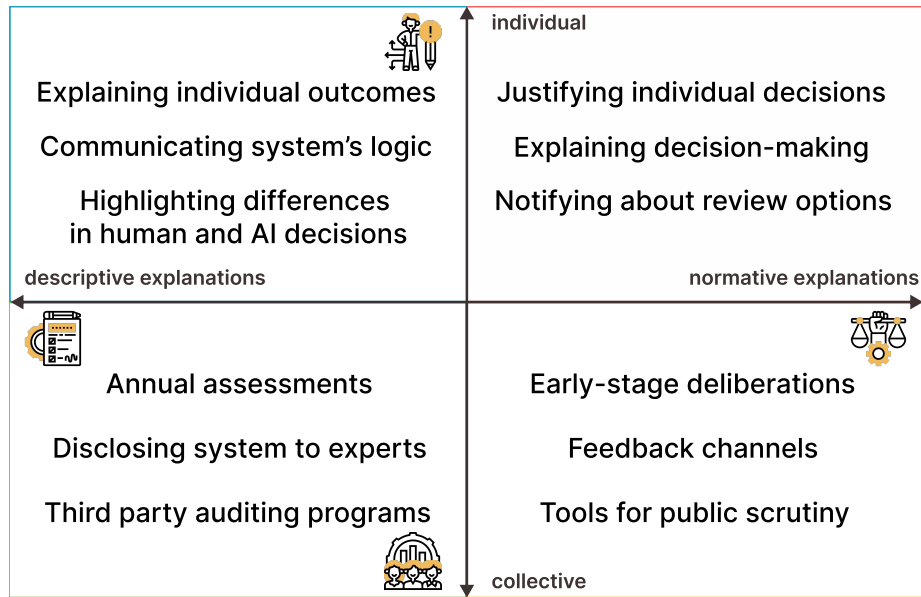


Fig. 6. **Duality of explainability and contestability.** The figure shows a grid created from two dimensions of explainability and contestability: descriptive versus normative explanations and individual versus collective action. Each quadrant contains examples of mechanisms that are assigned to the corresponding dimensions. We describe four examples: 'Justifying individual decisions' requires providing explanations to specific decision subjects (individual) in order to explain why a decision is correct and fair (normative). 'Early-stage deliberations' are likely focussed on aligning the values of an AI system with those of the citizens (normative), thus providing a form of collective democratic control. 'Third party auditing programs' aim to capture a detailed technical account of the AI system (descriptive) in collaboration with experts and the deploying institution (collective). Lastly, by 'communicating the system's logic' public servants might provide decision subjects (individual) with a general overview of the algorithmic decision-making process (descriptive). The combination of both dimensions in a grid is based on insights from the interviews and aims to offer a new lens to conceptualize the duality of mechanisms in explainability and contestability.