Start Using Justifications When Explaining AI Systems to Decision Subjects

Klára Kolářová[1] and Timothée Schmude[2](✉)

[1] ETH Zurich, Zurich, Switzerland
[2] Faculty of Computer Science, Doctoral School Computer Science, University of Vienna, Vienna, Austria
timothee.schmude@univie.ac.at

Abstract. Every AI system that makes decisions about people has stakeholders who are affected by its outcomes. These stakeholders, whom we call decision subjects, have a right to understand how their outcome was produced and to challenge it. Explanations should support this process by making the algorithmic system transparent and creating an understanding of its inner workings. However, we argue that while current explanation approaches focus on *descriptive* explanations, decision subjects also require *normative* explanations or *justifications*. In this position paper, we advocate for justifications as a key component in explanation approaches for decision subjects and make three claims to this end, namely that justifications i) fulfill decision subjects' information needs, ii) shape their intent to accept or contest decisions, and iii) encourage accountability considerations throughout the system's lifecycle. We propose four guiding principles for the design of justifications, provide two design examples, and close with directions for future work. With this paper, we aim to provoke thoughts on the role, value, and design of normative information in explainable AI for decision subjects.

Keywords: explainability · contestability · regulation · policy · AI · interdisciplinary research

1 Introduction

Explainability is seen as a cornerstone of trustworthy AI because it allows stakeholders to understand and assess algorithmic decisions. But some actions, like contestation [1], require more than transparency of the system; they require transparency of rationale [15]. Every algorithmic decision-making (ADM) system that is used to decide about humans has stakeholders who are affected by these outcomes – *decision subjects*. Their right to understand and contest decisions is granted by legal texts governing the development and use of AI systems

K. Kolářová and T. Schmude—These authors contributed equally to this work.

L. Hagedorn et al. (Eds.): DIGHUM 2025, LNCS 16319, pp. 190–202, 2026.
https://doi.org/10.1007/978-3-032-11108-1_14

in the EU [33], including the AI Act, GDPR, and DSA[1]. Although the realization of these rights is dependent on the concrete application domain, assessed risk, and operational context [6], the principles of explainability and contestability are embedded in the normative frameworks underlying these regulations and constitute a central step on the path towards trustworthy AI systems [43].

However, designing for these rights raises challenges: Deciding whether and how to contest requires an understanding of the system that made the decision [1]. But to *understand*, decision subjects require not only "mechanistic" information, such as how the model processes inputs and how features are weighed, but also "intentional" information, such as which purpose the system serves and how it is used. Humans interpret algorithmic systems as intentional agents, which creates an information need for normative reasoning [14,49]. Yet while information about intention and rationale is equally important in explaining algorithmic decisions to decision subjects [20], explainable AI (XAI) has so far mostly focused on providing the first, mechanistic type. The second type of information can be conveyed through *justifications*, which we define as describing the norms and values that guide a system's deployment [7,20], and in distinction to *generated* justifications that merely rationalize discrete model decisions [9].

In this position paper, we argue that explainable AI should incorporate justifications as a key informational component, especially when explaining algorithmic decisions to persons affected by them. We make three claims to this end: First, justifications address the epistemic needs of decision subjects by explaining a system's rationale, a key aspect for judging a system's legitimacy, that is not usually covered by XAI approaches. Second, justifications help decision subjects to assess the acceptability of their individual outcome and of the system as a whole, which supports the choice of adequate contestation channels. Third, justifications incentivize stakeholders throughout the system's lifecycle to consider social and legal accountability for their decisions, producing a more trustworthy system in the process.

In Sects. 2, 3 and 4, we elaborate on each claim and refer to current literature to support them. In Sect. 5, we propose four guiding design principles stating that justifications should be *normative*, *argumentative*, *challengeable*, and *relational* and provide two examples of such designs. With this paper, we aim to provoke scholars and designers of explainable AI to consider the importance of normative information for decision subjects. We argue that explainable AI should not only answer the question "What does the system do?" but also "Why is the system justified in doing this?"

2 Justifications Are Required to Meet Decision Subjects' Epistemic Needs

Humans Interpret Algorithmic Systems as Intentional Agents. People do not encounter automated decision-making systems (ADM) as disinterested calculators. Rather, they tend to interpret these systems as intentional agents –

[1] General Data Protection Regulation and Digital Services Act, respectively.

entities that act with goals and purposes. This is a deeply rooted cognitive inclination, called the "intentional stance" [14]. Classic psychological experiments have shown that humans readily attribute agency and intent even to simple geometric shapes, if they move in ways that appear purposeful [19]. This interpretive bias extends to modern technologies: people interact with robots, conversational agents, and other intelligent-seeming systems as if they were social beings. While they do not assign them sentience per se, people perceive these systems as "hand puppets"–projections of the social agents in the background [11]. In consequence, even less human-like decision-making systems are perceived as embodying the intentions, beliefs, and values of the deploying institutions [10]. This changes by which standards its decisions are evaluated. Decision subjects ask not only *how* a certain outcome was produced – which we call *descriptive* explanations – but also *why* this outcome is "right" [41] – which we call *normative* explanations or *justifications*. A main point of this paper is that descriptive explanations do not directly speak to questions of legitimacy or value alignment [9]. These questions require justifications that link the system's behavior to overarching goals, principles, or purposes, i.e., its *rationale* [15]. This is reflected in the concept of "goal-driven XAI," where agents are explained not just through statistical methods, but through frameworks that reconstruct their beliefs, desires, and intentions [41]. Such approaches recognize that when a system appears to act for a reason, it will be interpreted and assessed in normative terms.

Current XAI Tools Are Limited and Not Designed for Normativity. While explainability methods have gained significant traction in recent years, most remain focused on creating mechanistic transparency. These "data-driven" [41] XAI approaches aim to reconstruct how a system reached its output through techniques such as feature importance or saliency mapping [39]. While useful in many contexts, this paradigm comes with several limitations. First, the explanations produced are inherently selective, often highlighting only part of the model's internal logic, which opens the door to skewed portrayals or cherry-picked narratives [46]. Second, in complex models, their verification against the actual decision process is impossible due to the very opacity that creates the need for XAI in the first place [39]. The traditional descriptive methods may be well suited to domains in which rules are based on objective "ground truths" (like medical imaging systems) and aggregated statistical evaluations of performance will indeed be central to justified deployment. But they fall short when applied to systems that affect individual rights, opportunities, or access to public goods based on human-defined and disputable values.

Affected Persons Require Justifications. In cases where events or actions affect a person's life, mechanistic and statistical causality chains are typically of secondary interest [35]. Instead, Langer et al. 2020 [25] identify fairness, informed consent, and, crucially, morality and ethics as key epistemic desiderata of persons affected by automated decisions. These cannot be satisfied through technical explanation alone, because evaluative questions (What makes this decision a good one? Who does it benefit? Do I agree with it?...) require normative information, not just computational insight [18]. In fact, justification may be more important to decision subjects than explanation. Vredenburgh 2024 [48] points out that even in purely human-driven decision-making, people rarely care about the exact cognitive process behind a choice. Instead, they want to know whether it can be defended in terms of norms, values, and legitimate goals. Since these normative judgments create epistemic desiderata not satisfiable by descriptive explanations alone, we argue that providing information to support such judgments is a key role of human-centered explainable AI, as is the development of design approaches to deliver this information.

3 Justifications Shape Decision Subjects' Intent to Accept or Contest Decisions

Transparency in Rationale Discourages Algocratic Shortcuts Through ADM Systems. Public institutions, by definition and by design, have power over the individual as they enforce state rules and sanctions [38]. Delegation of power is a common practice in modern representative democracies. Yet rather than blind compliance, deference should be "informed" by being the result of scrutiny, deliberation, and reflection [24]. In ADM systems, informed deference is often bypassed by "algocratic shortcuts" [4], i.e., delegating decision-making to supposedly neutral or objective [30] algorithms instead of incorporating citizen deliberation. This is analogous to deference to experts ("epistocratic shortcuts"). To discourage such shortcuts, ADM systems should instead be framed from the view of *digital civics*, i.e., as technology that uses collaborative and relational approaches [12]. Justifications can support these approaches by creating transparency in rationale [15], which can raise the perceived legitimacy of an ADM system while also avoiding complete deference to them.

Understanding the Objectives of an ADM System Can Improve Acceptance of Unfavorable Decisions. Institutions necessarily need to make decisions that are unfavorable to some. Providing justifications of these decisions can help make even tough decisions easier to accept, if the person agrees with the ultimate values and objectives of the deploying institution [4]. Further, including the point of view of the deploying institution makes it possible to re-assess the justifiability of the system as a whole. This dual assessment points towards a distinction between *local* and *global* justifications analogous to the distinction made for explanation approaches [33]. While local justifications can help people judge the acceptability of their individual decisions, global justifications in combination with explanations can improve credibility and create a sense of reassurance despite an unfavorable outcome [27]. This way of rendering decisions understandable not only in a rational but also in an intentional sense might improve the feeling of relatedness [40] and the perception of procedural justice, which are essential to model the relation between state and individual [4].

Justifications Enable Contestation. Transparency in rationale enables contestability both before (ex-ante) and after (ex-post) decisions. As part of the preemptive measures that are favored by current AI policy to minimize the potential harms of high-risk ADM systems [37], justifications can be used to consult external stakeholders and the public to scrutinize the legal and ethical norms that drive subsequent ADM system development [20]. As part of ex-post measures that address and remedy realized harms [37], justifications can help decision subjects to decide on adequate contestation channels, e.g., whether to use judicial or non-judicial means to challenge decisions [43]. Both legal and design research state that understanding is a precondition for contestation, and see explainability and contestability as tightly linked [1,33]. For example, explanations should help users to "understand the relevant capacities and limitations" of high-risk AI systems (AI Act), the "precise purposes" of automated decisions in content moderation (DSA), and the "logic involved" in data processing and subsequent profiling (GDPR) to enable contestation and create empowerment [33]. But to fulfill its supporting role for contestation, explainability also needs to convey the intentions and goals of the deploying institution. As we established above, this information is not only descriptive but also normative, and can only be provided in the form of justifications. We thus see justifications as a key aspect of the broader principle of contestability-by-design [1] and as a building block for the design of contestation mechanisms.

4 Justifications Encourage Considering Social and Legal Accountability Throughout the System's Lifecycle

Disclosing Goals and Values Is Required to Establish Accountability. Accountability is seen as a core principle of trustworthy ADM systems and is closely tied to notions of fairness, auditability, and the minimizing of negative impact [21]. But assigning and assessing social and legal accountability requires more than merely providing insight into a model's internal mechanics, as "accountability requires justification and justification requires explanation. The form of each should determine the form of the others" [17]. Further, normative aspects, in particular the choice to employ a machine-learning-based decision tool, as well as the criteria it evaluates, are fully explainable and have significant implications on the final design, which constitutes a good reason to subject them to scrutiny [45]. This intuition is captured in the two-part framework for explanations, namely: a good explanation provides (1) (normative) information required to describe the decision in terms of its goals and intentions and (2) (descriptive) information treated as part of the record for backing up the adherence to this norm [45]. If the deploying institution fails to deliver records that align with their norms, they can be held responsible, thus establishing accountability.

All Types of ADM Systems Embed Human Decision-Making and Rationale that Should Be Open to Validation. Algorithmic decision-making systems exist in a grey space between data-driven models and goal-directed agents [41]. Though their interactional autonomy is limited, they are not mere smart calculators. They are deployed to advance human-defined objectives and inherit intentionality from their designers and deployers. Barocas et al. 2023 [5] define three categories of automated decision-making that implement human intentions: The first kind automates pre-existing decision-making rules and is thus a mere computational expression of these rules (such as software to determine welfare eligibility). The second kind learns to predict decisions based on data from past decisions and thus incorporates human-defined rules that were applied in the past (such as automated grading systems). The third kind derives decision-making rules from the data based on the definition of a target or goal to be optimized (such as predictive policing), which again was chosen through human intent. Therefore, all ADM systems implement value and policy choices made by humans and, in order to create accountability, should open these choices to validation and audit [53].

Justifications for Decision-Subjects Must Work Alongside Structural Accountability Mechanisms. While enabling contestation and agency through transparency is important, the responsibility for evaluating whether automated decisions are legitimate or correct should lie with governing institutions, not the governed individuals. Institutional safeguards against errors, bias, or illegitimate objectives are crucial, as individual contestations alone cannot resolve systemic issues [23]. Moreover, as exemplified by the GDPR cookie consent mechanism, placing responsibility on end-users (affected persons) without simple agency mechanisms can lead to disengagement and the waiver of one's rights. Explanations provided to decision subjects should therefore be designed as tools for individual oversight, enabling them to assess whether a system's outcome aligns with the institution's normative claims *in light of their personal context.* If it does not, decision subjects must be provided with simple, low-barrier channels to contest the outcome. For instance, if an organisation claims to use a race-neutral algorithm, but the decision appears racially biased, decision subjects should be able to trigger a review. In this model, explanations and justifications act as interfaces for critical discourse with the authority [20] and, when embedded within frameworks for systemic accountability and institutional guarantees of legal and ethical standards, explanations and justifications together enable informed deference [4].

5 Crafting Justifications

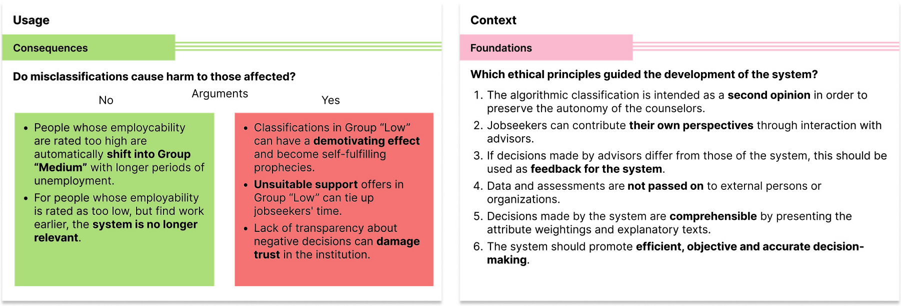
We established why justifications should be seen as a key component of explainable AI for decision subjects. In this Section, we propose guiding principles for the design of such justifications. With these principles, we aim to contribute a conceptual foundation that can be used to fit traditional explanation approaches to the rights and epistemic needs of decision subjects. The principles aim to introduce normative information (explaining goals, values, and intentions) to complement the mechanistic information (explaining features, outputs, and functions) in established explanation approaches. Namely, we propose that justifications in XAI should be:

- **normative** – describing the underlying norms, values, and intentions that guide an ADM system's actions as opposed to its mechanistic features, a value judgment rather than a technical assessment [26];
- **argumentative** – incorporating multiple points of view that could be advocated for by the involved stakeholders, supporting deliberation more than application, a forum rather than a manual [47];
- **challengeable** – presenting information in a way that invites reflection and opposition, justifying the reasoning behind a particular case, a debate rather than a factual authority [32];
- and **relational** – adapting to the relation between the sending and the receiving stakeholder [17], just as explanations are adapted to the knowledge and role of their audience.

The concrete design of justifications is context-dependent, as are most explanation approaches [16]. Designs can take the form of lists of pro and contra arguments, flowcharts of actor involvement in algorithmic decisions, or elaboration on the design choices that determined training data and model of a system.

To illustrate, we provide two examples for the design and content of justifications in Fig. 1. These justifications are based on information about an ADM system that was planned to be used in the Austrian employment agency "AMS" to rate the employability of jobseekers. The system's deployment was stopped in 2020 by Austrian Court of Administration, but comparable systems were actually implemented and are still in use in other European countries [44]. The justifications shown incorporate information from literature that documented the system's development and planned deployment [3,30], and follows a question-driven approach [29] to address two information needs of potential decision subjects [42]: *Can misclassifications harm people who are targeted by this system?* and *Which ethical principles guided the system's development?* The first question is addressed through an argumentative approach, presenting both information for and against the claim. The second question is addressed through a list of principles as given by the deploying institution [22]. The latter justification could be complemented by another sheet describing whether these principles are met in the actual deployment.

While not using the same terms to describe them, previous work has employed justifiability as a relevant dimension in explanation approaches. Lee et al. 2019 [27] co-designed an algorithmic food donation distribution system with stakeholders from a local community. In their design, affected stakeholders could choose which values to prioritize in the optimization (such as food access, income, and travel time). This direct participation in the design process increased people's acceptance even of decisions that favored another donation destination over their own, because it produced transparency in rationale and values. Since in resource distribution systems, the different values and priorities will be in tension and their priority subjective, extended designs have been proposed which allow decision subjects to contest the norms, reaching a system built on values



Usage

Consequences

Do misclassifications cause harm to those affected?

Arguments

No	Yes
• People whose employability are rated too high are automatically **shift into Group "Medium"** with longer periods of unemployment. • For people whose employability is rated as too low, but find work earlier, the **system is no longer relevant**.	• Classifications in Group "Low" can have a **demotivating effect** and become self-fulfilling prophecies. • **Unsuitable support** offers in Group "Low" can tie up jobseekers' time. • Lack of transparency about negative decisions can **damage trust** in the institution.

Context

Foundations

Which ethical principles guided the development of the system?

1. The algorithmic classification is intended as a **second opinion** in order to preserve the autonomy of the counselors.
2. Jobseekers can contribute **their own perspectives** through interaction with advisors.
3. If decisions made by advisors differ from those of the system, this should be used as **feedback for the system**.
4. Data and assessments are **not passed on** to external persons or organizations.
5. Decisions made by the system are **comprehensible** by presenting the attribute weightings and explanatory texts.
6. The system should promote **efficient, objective and accurate decision-making**.

Fig. 1. Two examples of justifications for decision subjects. The described system was intended to rate the employability of jobseekers to assist job counselors in the decision about adequate support measures [30]. For both examples, the leading questions cannot be answered easily through factual or descriptive information, but instead require considerations of the norms, values, and viewpoints involved. In this sense, they do not aim to provide a factually correct answer, but reasoning grounds for deliberation [34] on decisions, including whether to contest a decision and whether to contact someone in a supervisory position, such as a human in the loop [13]. Both examples adhere to the four proposed principles for justification design, namely, they are normative (based on norms and values), argumentative (give multiple views), challengeable (do not claim informational authority), and relational (are addressed towards decision subjects).

representing both the deployer and the affected stakeholders [20]. Further, Weitz et al. 2024 [50] conducted workshops with end-users from the public sector to identify their needs for XAI interfaces and found that social norms and cultural values were essential to assess whether an AI system aligned with their values.

Notably, justifications, just as explanations, come with their own set of limitations. Without the grounding evidence of explanations, justifications can be tuned to satisfy their reader without conveying truthful information about a system, which might in reality be discriminatory. Further, reasoning approaches are prone to "confirmation bias", as information confirming one's views is easier to accept than those to the contrary [47]. Hence, both approaches are necessary and complementary. While explanations document the factual, justifications provide a sense of guiding intention and purpose.

6 Future Work

Future work can follow several avenues to examine the design and effect of justifications in explainable AI: First, qualitative analyses could identify the information needs that decision subjects and other stakeholders have with respect to justifications specifically. Previous work established collections of information needs in the context of AI systems [28,42], but did not focus on the distinction between descriptive and normative information and their respective value to different stakeholders. Second, studying the role and effect of justifications in the context of contestation is a promising avenue for both qualitative and quantitative research. While contestation has been the subject of some empirical work [2,31,51], many open questions remain, such as which contestation options decision subjects prefer, which kind of information is most helpful to this end, and how contestation success can be measured. Third, while the design space for explanation has been explored in many studies [36], design for justifications – beyond those that justify individual algorithmic decisions with descriptive information [8,52] – is comparably unexplored. Finding ways to convey the challengeable norms and principles guiding an ADM system understandably and helpfully to decision subjects is thus another fruitful direction of future work.

7 Conclusion

In this paper, we argued that justifications should be a key component when explaining algorithmic decisions to decision subjects. We made three claims to this end: justifications fulfill key epistemic desiderata of decision subjects; justifications can shape decision subjects' intent to accept or contest unfavorable decisions; and justifications encourage consideration of accountability along the whole lifecycle of ADM systems. We provide two examples of such justifications for the use case of an algorithmic decision-making system in the employment domain. To translate these claims into practical approaches, future work should conduct requirement analyses to study decision subjects' information needs for justifications and develop designs that employ justifications in complement with explanations to be of value to decision subjects.

Acknowledgments. This work has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20058].

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Alfrink, K., Keller, I., Kortuem, G., Doorn, N.: Contestable ai by design: towards a framework. Mind. Mach. **33**(4), 613–639 (2023)
2. Alfrink, K., Keller, I., Semperena, M.Y., Bulygin, D., Kortuem, G., Doorn, N.: Envisioning contestability loops: evaluating the agonistic arena as a generative metaphor for public ai. She Ji: J. Design, Econ. Innovat. **10**(1), 53–93 (2024)
3. Allhutter, D., Mager, A., Cech, F., Fischer, F., Grill, G.: Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Tech. rep., Österreichische Akademie der Wissenschaften (2020), epub.oeaw.ac.at
4. Alnemr, N.: Democratic self-government and the algocratic shortcut: the democratic harms in algorithmic governance of society. Contemporary Political Theory **23**(2), 205–227 (2024)
5. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning: Limitations and Opportunities. MIT Press (2023)

6. Bell, A., Nov, O., Stoyanovich, J.: The algorithmic transparency playbook. Center for Responsible AI. Disponible en: https://dataresponsibly github. io/algorithmic-transparency-playbook/resources/transparency_playbook_camera_ready. pdf (Consultado: 7 de febrero de 2024) (2023)
7. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (Apr 2018). https://doi.org/10.1145/3173574.3173951
8. Biran, O., Cotton, C.V.: Explanation and justification in machine learning: a survey. In: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI) (2017). https://www.ijcai.org/proceedings/2017/0202.pdf
9. Biran, O., McKeown, K.: Human-centric justification of machine learning predictions. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 1461–1467 (2017). https://doi.org/10.24963/ijcai.2017/202
10. Byrne, R.M.: Good explanations in explainable artificial intelligence (XAI): evidence from human explanatory reasoning. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, pp. 6536–6544. Macau, SAR China (Aug 2023). https://doi.org/10.24963/ijcai.2023/733, https://www.ijcai.org/proceedings/2023/733
11. Clark, H.H., Fischer, K.: Social robots as depictions of social agents. Behav. Brain Sci. **46**, e21 (2023). https://doi.org/10.1017/S0140525X22000668
12. Crivellaro, C., et al.: Infrastructuring public service transformation: creating collaborative spaces between communities and institutions through HCI research. ACM Trans. Comput.-Hum. Interact. **26**(3), 1–29 (2019)
13. Crootof, R., Kaminski, M.E., Ii, W.N.P.: Humans in the loop. Vanderbilt Law Rev. **76** (2023). https://doi.org/10.2139/ssrn.4066781
14. Dennett, D.C.: The intentional stance. A Bradford book, MIT Press, Cambridge, Mass., 7. printing edn. (1998)
15. de Fine Licht, K., de Fine Licht, J.: Artificial intelligence, transparency, and public decision-making. AI & Soc. **35**(4), 917–926 (2020). https://doi.org/10.1007/s00146-020-00960-w
16. Freiesleben, T., König, G.: Dear XAI community, we need to talk! In: Longo, L. (ed.) Explainable Artificial Intelligence, pp. 48–65. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-44064-9_3
17. Gillis, T.B., Simons, J.: Explanation < justification: Gdpr and the perils of privacy. J. Law & Innovation **2**, 71 (2019). https://scholarship.law.columbia.edu/faculty_scholarship/3132
18. Gold, N., Colman, A., Pulford, B.: Normative theory in decision making and moral reasoning. Behav. Brain Sci. **34**, 256–7 (2011). https://doi.org/10.1017/S0140525X11000495
19. Heider, F., Simmel, M.: An experimental study of apparent behavior. Am. J. Psychol. **57**(2), 243–259 (1944), http://www.jstor.org/stable/1416950
20. Henin, C., Le Métayer, D.: Beyond explainability: justifiability and contestability of algorithmic decision systems. AI & Soc. **37**(4), 1397–1410 (2022)
21. High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI (Apr 2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-AI
22. Holl, J., Kernbeiß, G., Wagner-Pinter, M.: Personenbezogene wahrscheinlichkeitsaussagen (»algorithmen«): Stichworte zur sozialverträglichkeit. Technical concept, Synthesis Forschung, Vienna (2019)

23. Kaminski, M.E.: Binary governance: lessons from the gdpr's approach to algorithmic accountability. Southern California Law Rev. **92**(6), 1529–1616 (2019). https://scholar.law.colorado.edu/faculty-articles/1265/, Accessed 11 Jul 2025
24. Lafont, C.: Democracy without shortcuts. Constellations **26**(3), 355–360 (2019). https://doi.org/10.1111/1467-8675.12432
25. Langer, M., et al.: What do we want from explainable artificial intelligence (XAI)? - a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif. Intell. **296**, 103473 (2021)
26. Langley, P.: Explainable, normative, and justified agency. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33(01), pp. 9775–9779 (2019)
27. Lee, M.K., et al.: WeBuildAI: participatory framework for algorithmic governance. Proc. ACM Hum.-Comput. Interact. **3**(CSCW), 1–35 (2019). https://doi.org/10.1145/3359283, https://dl.acm.org/doi/10.1145/3359283
28. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020, pp. 1–15. Association for Computing Machinery, New York (2020).https://doi.org/10.1145/3313831.3376590, https://doi-org.uaccess.univie.ac.at/10.1145/3313831.3376590
29. Liao, Q.V., Pribić, M., Han, J., Miller, S., Sow, D.: Question-driven design process for explainable AI user experiences (2021)
30. Lopez, P.: Reinforcing intersectional inequality via the AMS algorithm in Austria. In: Proceedings of the 18th Annual STS Conference, pp. 289–309. Graz (2019). https://doi.org/10.3217/978-3-85125-668-0-16
31. Lyons, H., Miller, T., Velloso, E.: Algorithmic decisions, desire for control, and the preference for human review over algorithmic review. In: 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 764–774. ACM, Chicago IL USA (Jun 2023). https://doi.org/10.1145/3593013.3594041, https://dl.acm.org/doi/10.1145/3593013.3594041
32. Lyons, H., Velloso, E., Miller, T.: Designing for Contestation: Insights from Administrative Law (Feb 2021). http://arxiv.org/abs/2102.04559
33. Maxwell, W., Dumas, B.: Meaningful XAI based on user-centric design methodology: Combining legal and human-computer interaction (HCI) approaches to achieve meaningful algorithmic explainability. SSRN Electron. J. (2023)
34. Mercier, H., Sperber, D.: Why do humans reason? Arguments for an argumentative theory. Behav. Brain Sci. **34**(2), 57–74 (2011). https://doi.org/10.1017/S0140525X10000968
35. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
36. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems (2020). https://arxiv.org/abs/1811.11839
37. Pi, Y., Proctor, M.: Toward empowering AI governance with redress mechanisms. Cambridge Forum on AI: Law Governance **1** (2025). https://doi.org/10.1017/cfl.2025.9
38. Prifti, K., Morley, J., Novelli, C., Floridi, L.: Regulation by design: features, practices, limitations, and governance implications. Mind. Mach. **34**(2), 13 (2024)
39. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
40. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. (2000)

41. Sado, F., Loo, C.K., Liew, W.S., Kerzel, M., Wermter, S.: Explainable goal-driven agents and robots - a comprehensive review. ACM Comput. Surv. **55**(10) (2023). https://doi.org/10.1145/3564240
42. Schmude, T., Koesten, L., Möller, T., Tschiatschek, S.: Information that matters: exploring information needs of people affected by algorithmic decisions. Int. J. Hum. Comput. Stud. **193**, 103380 (2025). https://doi.org/10.1016/j.ijhcs.2024.103380
43. Schmude, T., Yurrita, M., Alfrink, K., Goff, T.L., Viard, T.: Two Means to an End Goal: Connecting Explainability and Contestability in the Regulation of Public Sector AI (2025). https://arxiv.org/abs/2504.18236
44. Scott, K.M., Wang, S.M., Miceli, M., Delobelle, P., Sztandar-Sztanderska, K., Berendt, B.: Algorithmic tools in public employment services: towards a jobseeker-centric perspective. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2022, pp. 2138–2148. ACM, Seoul Republic of Korea (Jun 2022). https://doi.org/10.1145/3531146.3534631, https://dl.acm.org/doi/10.1145/3531146.3534631
45. Strandburg, K.J.: Rulemaking and inscrutable automated decision tools. Columbia Law Rev. **119**(7), 1851–1886 (2019), also available as NYU Public Law Research Paper No. 20-36
46. Sullivan, E.: Sides: Separating idealization from deceptive 'explanations' in xai. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, pp. 1714–1724. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3630106.3658999
47. Vassiliades, A., Bassiliades, N., Patkos, T.: Argumentation and explainable artificial intelligence: a survey. Knowl. Eng. Rev. **36** (2021). https://doi.org/10.1017/s0269888921000011
48. Vredenburgh, K.: Transparency and explainability for public policy. LSE Public Policy Rev. **3**(3) (2024)
49. Waytz, A., Morewedge, C.K., Epley, N., Monteleone, G., Gao, J.H., Cacioppo, J.T.: Making sense by making sentient: effectance motivation increases anthropomorphism. J. Pers. Soc. Psychol. **99**(3), 410 (2010)
50. Weitz, K., Schlagowski, R., André, E., Männiste, M., George, C.: Explaining it your way - findings from a co-creative design workshop on designing xai applications with ai end-users from the public sector. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Association for Computing Machinery, New York (2024).https://doi.org/10.1145/3613904.3642563
51. Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N., Bozzon, A.: 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, pp. 1–21. ACM, Hamburg Germany (Apr 2023). https://doi.org/10.1145/3544548.3581161, https://dl.acm.org/doi/10.1145/3544548.3581161
52. Zhou, J., Joachims, T.: How to explain and justify almost any decision: potential pitfalls for accountability in AI decision-making. In: 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 12–21. ACM, Chicago IL USA (Jun 2023). https://doi.org/10.1145/3593013.3593972, https://dl.acm.org/doi/10.1145/3593013.3593972
53. Züger, T., Asghari, H.: AI for the public. how public interest theory shifts the discourse on AI. AI & Soc. **38**(2), 815–828 (2023). https://doi.org/10.1007/s00146-022-01480-5