Contestable AI by Design

Towards a Framework

Alfrink, Kars; Keller, A.I.; Kortuem, G.W.; Doorn, N.

DOI
10.1007/s11023-022-09611-z

Publication date
2022

Document Version
Final published version

Published in
Minds and Machines: journal for artificial intelligence, philosophy and cognitive sciences

Citation (APA)
Alfrink, K., Keller, A. I., Kortuem, G. W., & Doorn, N. (2022). Contestable AI by Design: Towards a Framework. *Minds and Machines: journal for artificial intelligence, philosophy and cognitive sciences*, *33*(4), 613-639. https://doi.org/10.1007/s11023-022-09611-z

Minds and Machines
https://doi.org/10.1007/s11023-022-09611-z

Contestable AI by Design: Towards a Framework

Kars Alfrink[1] · **Ianus Keller**[2] · **Gerd Kortuem**[1] · **Neelke Doorn**[3]

Received: 21 August 2021 / Accepted: 4 August 2022

Abstract
As the use of AI systems continues to increase, so do concerns over their lack of fairness, legitimacy and accountability. Such harmful automated decision-making can be guarded against by ensuring AI systems are contestable by design: responsive to human intervention throughout the system lifecycle. Contestable AI by design is a small but growing field of research. However, most available knowledge requires a significant amount of translation to be applicable in practice. A proven way of conveying intermediate-level, generative design knowledge is in the form of frameworks. In this article we use qualitative-interpretative methods and visual mapping techniques to extract from the literature sociotechnical features and practices that contribute to contestable AI, and synthesize these into a design framework.

Keywords Artificial intelligence · Automated decision-making · Contestability · Design · Human–computer interaction · Machine learning · Sociotechnical systems

1 Introduction

Artificial Intelligence (AI) systems are increasingly used to make automated decisions that impact people to a significant extent. As the use of AI for automated decision-making increases, so do concerns over its harmful social consequences,

✉ Kars Alfrink
c.p.alfrink@tudelft.nl

Ianus Keller
a.i.keller@tudelft.nl

Gerd Kortuem
g.w.kortuem@tudelft.nl

Neelke Doorn
n.doorn@tudelft.nl

[1] Sustainable Design Engineering, TU Delft, Landbergstraat 15, 2628 CE Delft, The Netherlands

[2] Human Centered Design, TU Delft, Landbergstraat 15, 2628 CE Delft, The Netherlands

[3] Values, Technology and Innovation, TU Delft, Jaffalaan 5, 2628 BX Delft, The Netherlands

Published online: 13 August 2022

including the undermining of democratic rule of law and the infringement of basic human rights to dignity and self-determination (e.g. Chiusi et al., 2020; Crawford et al., 2019). A way to counteract such harmful automated decision-making is through *contestability*. Contestable AI systems are open and responsive to human intervention throughout their lifecycle: not only after an automated decision has been made, but also during its design and development.

A small but growing body of research explores the concept of contestable AI (Almada, 2019; Henin & Le Métayer, 2021; Hirsch et al., 2017; Lyons et al., 2021; Sarra, 2020; Vaccaro et al., 2019, 2021). However, although many do make practical recommendations, very little of this research is presented in a format readily usable in design practice. One such form of "intermediate-level generative design knowledge" (Höök & Löwgren, 2012; Löwgren et al., 2013) are *design frameworks*.

In this contribution we use qualitative interpretative methods supported by visual mapping techniques to develop a preliminary design framework that synthesizes elements identified through a systematic literature review, that contribute to contestability of AI systems. This preliminary framework serves as a starting point for subsequent testing and validation in specific application contexts.

Our framework consists of five system features and six development practices that contribute to contestable AI. The features are: 1. built-in safeguards against harmful behavior; 2. interactive control over automated decisions; 3. explanations of system behavior; 4. human review and intervention requests; and 5. tools for scrutiny by subjects or third parties. The practices are: 1. ex-ante safeguards; 2. agonistic approaches to machine learning (ML) development; 3. quality assurance during development; 4. quality assurance after deployment; 5. risk mitigation strategies; and 6. third-party oversight. We also offer a diagram for each set, capturing how features relate to various actors in a typical AI system, and how practices relate to typical AI system lifecycle stages.

This paper is structured as follows: First we discuss why contestability is a necessary quality of AI systems used for automated decision-making. Then we situate our efforts in the larger field of responsible design for AI. We subsequently frame design frameworks as generative, intermediate-level knowledge. We then describe our method of constructing the design framework. Following this, we describe the literature review, and the elements we have identified in the included sources. Finally, we discuss the synthesis of these elements into our proposed design framework. We end with some concluding remarks.

2 Contestability in Automated Decision-Making

The main focus of our effort is to ensure AI systems are open and responsive to contestation by those people directly or indirectly impacted throughout the system lifecycle. We define AI broadly, following Suchman (2018): "[a] cover term for a range of techniques for data analysis and processing, the relevant parameters of which can be adjusted according to either internally or externally generated feedback".

A growing number of scholars argue for contestability of AI systems in general, and in automated decision-making specifically (Almada, 2019; Hirsch et al., 2017; Sarra, 2020; Vaccaro et al., 2019).

Hirsch et al. (2017) describe contestability as "humans challenging machine predictions". They claim models are and will continue to be fallible. In many cases, the cost of "getting it wrong" can be quite high for decision subjects, and those human controllers held responsible for AI system performance. Contestability ensures such failures are avoided by allowing human controllers to intervene before machine decisions are put into force.

Vaccaro et al. (2019) argue that contestability can surface values, align design practice with context of use, and increase the perceived legitimacy of AI systems. Contestability is a "deep system property", representing a coming together of human and machine to jointly make decisions. It aids iteration on decision-making processes and can be aimed at human controllers ("experts") but also decision subjects. Contestability is a form of procedural justice, a way of giving voice to decision subjects, which increases perceptions of fairness, in particular for marginalized or disempowered populations.

Almada (2019) argues that contestability protects decision subjects against flawed machine predictions, by enabling *human intervention*. Such human intervention can take place not only post-hoc, in response to an individual decision, but also ex-ante, as part of AI system development processes (Kamarinou et al., 2016). Ex-ante contestability allows for an "agonistic debate", both internal and external, about data and modeling choices made to represent decision subjects, ensuring decisions comply with scientific, legal and democratic standards and values (Hildebrandt, 2017). Thus, contestability protects human self-determination and ensures human control over automated systems. Significant decisions do not only happen once a system is in operation and acting on subjects. Decisions are made throughout the system lifecycle. Contestability should therefore be part of the entire AI system development process: the practice of "contestability by design".

Finally, for Sarra (2020) contestability includes, but also exceeds, mere human intervention. Furthermore, it is distinct from simple opposition to automated decision-making. Instead, to contest is to engage with the substance of decisions *themselves*. It is more than voicing ones opinion. It requires an "articulate act of defense". Such a defense requires arguments, and arguments need information. In this case, an explanation of the decision made. This must include both a description of the "how" and a justification of the "why". Therefore, contestability demands explainability, and insofar as such explanations must include a *justification* specific to the case at hand, contestability also increases accountability. Most notably, contestability requires a "procedural relationship". A "human in the loop" is insufficient if there is no possibility of a "dialectical exchange" between decision subject and human controller. Without such dialogue, there can be no exchange of arguments specific to the case at hand.

In summary, contestability helps to protect against fallible, unaccountable, illegitimate, and unjust automated decision-making, by ensuring the possibility of human intervention as part of a procedural relationship between decision subjects and human controllers. The aim of this contribution is to develop a proposal for a

framework for contestability both as an AI system quality (contestability features), and an AI system development practice ("contestability by design").

3 Responsible Design for AI

As the adoption of AI continues to increase, so do concerns over its shortcomings, including lack of fairness, legitimacy and accountability. Such concerns cannot be met by purely technical solutions. They require a consideration of social and technical aspects in conjunction. This sociotechnical view emphasizes technical and social dimensions are entangled, producing specific outcomes irreducible to constitutive components (Franssen, 2015; Kroes et al., 2006). What is more, AI systems are distinct from "traditional" sociotechnical systems because they include "artificial agents" and humans interacting in a dynamic evolving environment (van de Poel, 2020). As a result, AI systems contain a particularly high degree of uncertainty and unpredictability.

Design, human–computer interaction (HCI) design in particular, is uniquely equipped to tackle such sociotechnical challenges, because it draws on both computer science and social science, joining positivist and interpretive traditions (Dourish, 2004; Katell et al., 2020; Tonkinwise, 2016). This allows interaction design to more adequately "see" AI systems. By virtue of its roots in traditional design, HCI design has the capacity to *act* in the face of complexity and ambiguity, by co-evolving problem and solution space in tandem (Dorst & Cross, 2001; Norman & Stappers, 2015).

However, current design knowledge aimed at "responsible" and "ethical" AI is often of a high level of abstraction, and not connected to specific application domains. A lot of work is left for designers to translate such knowledge to their own practice. To illustrate this point we briefly summarize a number of prominent systematic reviews and meta-analyses drawn from across disciplines (Jobin et al., 2019; Morley et al., 2019; Shneiderman, 2020).

Jobin et al. (2019) identify eleven overarching ethical values and principles. These are, in order of frequency of the number of sources featuring them:

> transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity.

The first five principles are mentioned in over half of the sources. Importantly, Jobin et al note that, although there is convergence on the level of principles, the sources surveyed do diverge significantly in: 1. how they are interpreted; 2. why they are considered important; 3. what they should be applied to; and 4. how they should be implemented.

Morley et al. (2019) offer a more condensed set of themes, which together "define" ethically-aligned ML as:

> (a) beneficial to, and respectful of, people and the environment (beneficence); (b) robust and secure (non-maleficence); (c) respectful of human values

> (autonomy); (d) fair (justice); and (e) explainable, accountable and understandable (explicability).

Morley et al argue that principles are insufficient for changing actual AI systems design, and ethics scholars must do the hard work of translating the "what" of principles into the "how" of practices. By mapping principles to AI system lifecycle phases, they show current efforts are unevenly distributed, and where coverage exists, available solutions lack variety.

Finally, Shneiderman (2020) also notes there is a gap between principles and practice when it comes to "human-centered AI". They offer 15 recommendations organized in a "a three-layer governance structure:"

> (1) reliable systems based on sound software engineering practices, (2) safety culture through proven business management strategies, and (3) trustworthy certification by independent oversight.

Shneiderman also points out it is necessary to move beyond general statements, towards support for specific social practices.

In short, currently available knowledge related to responsible and ethical AI is often of a high level of abstraction. Furthermore, scholars surveying the field agree it is necessary to translate principles into practices. Our aim is therefore to create knowledge of a more intermediate level, situated between theory and specific instances, in the form of a design framework.

We focus on the principle of contestability in the context of automated decision-making. This principle stresses the sociotechnical character of AI systems: Contestability is about humans challenging machine decisions. It helps to surface values embedded in AI systems, aligning design with context of use. Contestability is a *deep system property*, linking humans and machines in joint decision-making. It enables *agonistic debate* about how models are made to represent the world in a particular way. Because human and AI decisions happen throughout the system lifecycle, what is needed is *contestability by design*.

In this paper we take the first step towards a design framework for contestable AI by summarizing ideas and mechanisms collated from previous work. Such mechanisms should align with the sociotechnical view, taking into account AI systems' entangled and volatile nature. Future efforts may then make ready use of the resulting provisional framework, for purposes of testing and validation in specific application contexts.

4 Design Frameworks as Generative Intermediate-Level Design Knowledge

We seek to construct a framework for the design of contestable AI systems. We conceive of a design framework as a form of "generative intermediate-level design knowledge" (Löwgren et al., 2013). *Generative* means it offers the seed for a design solution with particular qualities without fully prescribing its shape. *Intermediate-level* means it occupies a space between specific instances of designed artifacts, and

Table 1 Search terms used

Concept	Search terms used
Contestability	contestation (`contest*`), controversy (`controvers*`), debate (`debat*`), disagreement (`disagree*`), disputation, dispute (`disput*`), dissension (also dissention), dissensus (`dissen*`)
Artifical intelligence	artificial intelligence (also AI), machine learning (also ML), algorithmic system (`algorithm*`), automated decision-making
Design	design

generalized knowledge (theory). The design knowledge we seek to create describes particular sociotechnical system properties operationalizing the principle of contestability. We ground our framework in current knowledge on contestable AI. The purpose of the framework is to aid in the creation of designed artifacts. Following Stolterman and Wiberg (2010), we understand such design artifacts to be either in the service of improving a use situation, or in service of embodying new ideas (concepts) and theories. Our definition of "design framework" is aligned with Obrenović (2011). It should describe "the characteristics that a design solution should have to achieve a particular set of goals in a particular context", where our goal is contestable AI in the context of automated decision-making.

5 Method of Design Framework Construction

We performed the following steps to construct our framework: We used a systematic review to collect sources discussing contestable AI. We then used reflexive thematic analysis to construct from the literature a number of elements contributing to contestable AI. Finally, we used visual mapping techniques to synthesize these elements into framework diagrams.

5.1 Data Collection

Our data-collection procedure broadly follows Moher et al. (2009). Using Scopus, we searched for journal articles and conference papers published between 2016 and 2021 mentioning in their title, abstract or keywords "AI", "contestability" and "design". Synonyms for contestability were selected from the Merriam-Webster thesaurus entry for "contestation"[1]. We used our best judgment to decide on related

[1] Merriam-Webster. (n.d.). Contestation. In Merriam-Webster.com thesaurus. Retrieved May 28, 2021, from https://www.merriam-webster.com/thesaurus/contestation

terms for AI. See Table 1 for an overview of search terms used. The exact Scopus search is as follows:

```
TITLE-ABS-KEY( (design*) AND (contest* OR contro-
vers* OR debat* OR disagree* OR disput* OR dissen*)
AND ("artificial intelligence" OR "AI" OR "machine
learning" OR "ML" OR algorithm* OR "automated deci-
sion making") ) AND (PUBYEAR> 2015) AND (PUBYEAR<
2022) AND (LIMIT-TO(DOCTYPE, "cp") OR LIMIT-TO (DOC-
TYPE, "ar"))
```

We collated the results, and first removed duplicates. Then, using Rayyan (Ouzzani et al., 2016), we manually screened records' titles and abstracts for actually referring to contestability (rather than e.g. "contest" in the sense of a competition). The resulting set was assessed for eligibility on the basis of the full text. Here our criterion was whether papers did indeed discuss actionable sociotechnical system properties contributing to contestability. Once an initial set of inclusions was identified, we used Scopus to also screen 1. their references (i.e. "backward snowball"), and 2. all items referring to our inclusions (i.e. "forward snowball"). The resulting inclusions were once again assessed for eligibility. We then performed one final round of snowballing, screening, and qualitative assessment on the new inclusions. Figure 1 shows the stages of our systematic review.

5.2 Analysis & Synthesis

Our approach to analysis and synthesis is adapted from reflexive thematic analysis as described by Braun and Clarke (2006). Our procedure was as follows: Analysis was done in Atlas.ti (version 22 on MacOS). We read the included sources and selected those passages discussing what we might call "active ingredients": actionable sociotechnical system properties contributing to contestability. We grouped similar passages together, and assigned a label to each grouping capturing the essence of the property it represents. We then took the resulting list of properties, and looked for hierarchical and lateral relationships. In this step we relied heavily on visual mapping techniques, and used existing diagrams as a foundation. Once we had our preliminary framework, we checked the result against the selected passages, and against an end-to-end read-through of the source literature, to verify the framework properly covers and reflects it.

6 Elements in Extant Literature Contributing to Contestable AI

This section describes the elements we have identified in the literature. We have categorized them as either features or practices. They are summarized in Tables 2, 3 and 4, and are described in detail in the following sections.

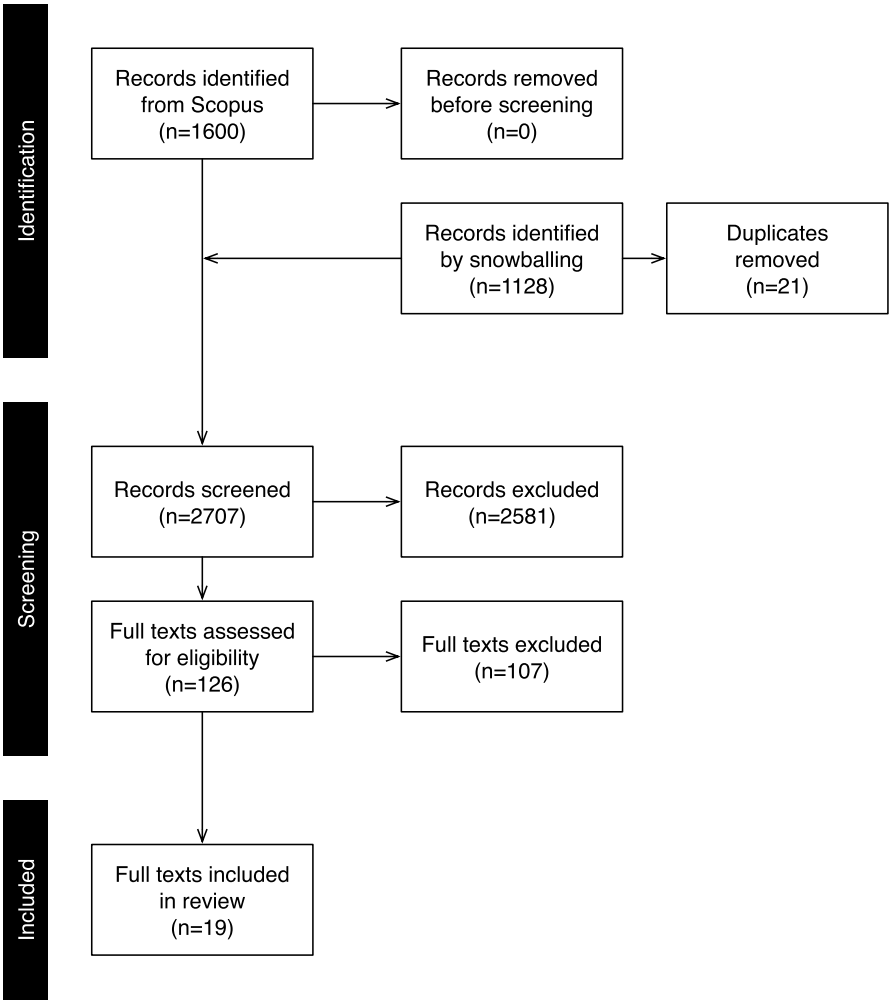

Fig. 1 Flow of information through the different phases of the systematic review

6.1 Features

6.1.1 Built-in Safeguards Against Harmful Behavior

This feature introduces procedural safeguards limiting what AI systems can do unilaterally. One such safeguard is to make the automated decision-making process *itself* adversarial. This can be achieved by introducing a second automated system external to the controlling organization, which machine decisions are run through. If disagreement between both systems occurs, decision can be flagged for human review, or automated dispute resolution mechanisms can take over. Such adversarial procedures could occur on an ongoing basis, or at the request of human controllers

or decision subjects. An additional benefit of a second (possibly public) system that decisions need to pass through is the creation of a record of all decisions made, which can aid outside scrutiny (Almada, 2019; Elkin-Koren, 2020; Lyons et al., 2021).

In some cases, it may be necessary and possible to implement formal constraints on system behavior. These would protect against undesired actions, and demonstrate compliance with standards and legislation (Aler Tubella et al., 2020).

6.1.2 Interactive Control Over Automated Decisions

This feature is primarily aimed at human controllers, although in some cases it may also be made available to decision subjects. It enables direct intervention in machine decisions. In HCI, the concept of *mixed-initiative interaction* refers to shared control between intelligent systems and system users. Such an approach may also be employed in the case of decision-support or semi-automated decisions. The final decision would be the result of a "negotiation" between system and user (Kluttz & Mulligan, 2019; Novick & Sutton, 1997 in Vaccaro et al., 2019) In some cases it may be possible to allow users to correct or override a system decision. This is of particular importance in a decision-support setting, where such corrections may also function as a feedback loop for further system learning (Bayamlıoğlu, 2021; Hirsch et al., 2017; Vaccaro et al., 2019, 2020). Where direct override is not a possibility, some form of control can be offered in an indirect manner by allowing users to supplement the data a decision is based on with additional contextual information (Hirsch et al., 2017; Jewell, 2018).

6.1.3 Explanations of System Behavior

This feature is primarily aimed at decision subjects but can also be of use to human controllers. It helps actors understand the decisions made by AI systems. A decision subject should know a decision has been made, that there is a means of contesting, and be provided with an explanation of the decision (Lyons et al., 2021). Explanations should contain the information necessary for a decision subject to exercise their rights to human intervention and contestation (Bayamlıoğlu, 2021; Lyons et al., 2021; Ploug & Holm, 2020).

Individual decisions should be reproducible and *traceable*. It should be possible to verify the compliance of individual decisions with norms. This requires version control, and thorough record-keeping (Aler Tubella et al., 2020). Simply keeping an internal log could already be a huge improvement. These records should include the state of the model, the inputs, and decision rules at the time of producing a specific outcome (Bayamlıoğlu, 2021). The norms decisions should adhere to should be elicited and specified ex ante (Aler Tubella et al., 2020).

Explanations should not simply be a technical account of how a model's output relates to its input. It should also include the organizational, social and legal context of the decision. In other words, the emphasis shifts from explaining the computational rules to the decision rules, offering a *behavioral model* of the AI system as a whole, from a sociotechnical perspective (Aler Tubella et al., 2020; Almada,

2019; Brkan, 2019; Crawford, 2016; Hirsch et al., 2017). This behavioral approach accounts for the limitations of transparency efforts focusing on "the algorithm" in isolation (Ananny & Crawford, 2018 in Henin & Le Métayer, 2021). It also seeks to strike a balance between usability and comprehensiveness, in an effort to avoid the "transparency paradox" (Nissenbaum, 2011 in Crawford, 2016).

These requirements should be satisfiable even for models that are opaque due to their technical nature. Nevertheless, it may be desirable to reduce model complexity, e.g. by limiting the number of features under consideration, or by using fundamentally more intelligible methods (e.g. decision trees vs. deep neural networks) (Bayamlıoğlu, 2021).

Although explanations may be of a static form, if deep understanding and exploration of counterfactual scenarios is desired, "sandboxing" or "black box in a glass box" approaches are worth considering. Using these approaches, users are able to manipulate inputs and see how these affect outputs. These techniques can work without needing to fully describe decision rules, which may be useful for cases where these cannot or will not be disclosed (Höök et al., 1998 in Hirsch et al., 2017). By offering explanations that include confidence levels, human controllers can direct their focus to those decisions warranting closer scrutiny (Hirsch et al., 2017; Vaccaro et al., 2019).

Another way to deal with model opacity (due to their proprietary or sensitive nature) is to generate local approximations using techniques such as "model inversion". However, once again we emphasize not to fixate on the technical components of AI systems in isolation (Hirsch et al., 2017; Leahu, 2016; Mahendran & Vedaldi, 2015; Ribeiro et al., 2016; Tickle et al., 1998 in Edwards & Veale, 2018).

Explanations in the service of contestability should not simply describe why a decision was made, but also why the decision is considered *good*. In other words, decision subjects should receive a *justification* as well. This avoids the self-production of norms (Rouvroy, 2012 in Henin & Le Métayer, 2021).

6.1.4 Human Review and Intervention Requests

This feature is aimed at decision subjects, and third parties acting on behalf of decision subject individuals and groups. It gives subjects the ability to "ask questions and record disagreements", both on the individual and the aggregate scale (Hirsch et al., 2017; Ploug & Holm, 2020; Vaccaro et al., 2019).

Human controllers and decision subjects should not be mere passive recipients of automated decisions. They should be put in dialogue with AI systems. Reliance on out-of-system mechanisms for contestation is insufficient (Kluttz et al., 2019 in Henin & Le Métayer, 2021).

A commonly recommended mechanism for responding to post-hoc contestation is human review and intervention (Lyons et al., 2021). Requests for human intervention are necessarily post-hoc, since they happen in response to discrete decisions, when a subject feels a decision has harmed or otherwise impacted their rights, freedoms or interests (Almada, 2019). Such intervention requests could be facilitated through auxiliary platforms, or be part of the system itself (Almada, 2019; Bayamlıoğlu, 2021). Although existing internal or external review procedures are

sometimes considered sufficient, in many cases new mechanisms for contestation will be required. Due process mechanisms should be designed into the AI systems itself (Lyons et al., 2021).

Human review is seen as an antidote to machine error. Human controllers can use tacit knowledge, intuition, and access to contextual information to identify and correct harmful automated decisions. In this way, allowing for human intervention is a form of quality control (Almada, 2019; Walmsley, 2021).

In the context of GDPR the right to human intervention is tied to fully automated decision-making only (Brkan, 2019). In practice, such a distinction may not be so clear-cut. From a sociotechnical perspective humans are always part of the decision chain leading up to a machine decision, in the role of designers, developers and operators. Furthermore, the mere presence of a human at the very end of the chain (the so-called "human in the loop") may not be a sufficient safeguard against machine error if human controllers do not have the authority or ability to base their final decision on more information than what was provided to them by the AI system (Almada, 2019). By extension, human controllers who respond to intervention request should have the authority and capability to actually *change* previous decisions (Brkan, 2019).

It is of course entirely possible for human intervention to be biased, leading to worse outcomes compared to a fully automated decision. This should be guarded against by introducing comparative measures of the performance of human-controlled and fully automated procedures (Almada, 2019). AI system controllers must make room within their organizations for receiving, evaluating and responding to disputes (Sarra, 2020).

Channels for contestation should be clear, accessible, affordable and efficient so that further harm to subjects is minimized (Lyons et al., 2021; Vaccaro et al., 2021). Mechanisms for requesting human intervention should provide "scaffolding for learning" (Applebee & Langer, 1983; Salehi et al., 2017 in Vaccaro et al., 2020). Documentation of the decision-making procedures should be integrated with the appeal procedure and communicated in alternative formats to ease comprehension (Vaccaro et al., 2020) and to help subjects in formulating their argument (Lyons et al., 2021; Vaccaro et al., 2021)

A risk of appeal procedures is that burdens are shifted to individual subjects. Ways of addressing this include allowing for synchronous communication with decision makers (Vaccaro et al., 2021), or to have third parties represent subjects (Bayamlıoğlu, 2021; Edwards & Veale, 2018; Lyons et al., 2021; Vaccaro et al., 2020).

Another limitation of current appeal procedures is that they handle decisions individually (Vaccaro et al., 2019). Groups should be able to acquire explanations of decisions collectively. Developers should not only consider individual impacts, but also group impacts (Edwards & Veale, 2018). Mechanisms for contestability should allow for collective action, because harms can be connected to group membership (Lyons et al., 2021). One way to aid collective action would be to publicize individual appeals cases so subjects can compare their treatment to those of others, and identify fellow sufferers (Matias et al., 2015; Myers West 2018; Sandvig et al., 2014

in Vaccaro et al., 2020). Subjects should be supported in connecting to those who share their fate (Vaccaro et al., 2021).

Any kind of human intervention in response to decision subjects' appeals may not qualify as actual contestation. Decision subjects should be able to express their point of view, if only to provide additional information based on which a decision may be reconsidered (Bayamlıoğlu, 2021). For true contestation to be the case, not only should the subject be allowed to express their point of view, but there should also be a *dialectical exchange* between subject and controller (Mendoza & Bygrave, 2017 in Brkan 2019). Therefore, contestation includes human intervention, but should not be *reduced* to it. Care should also be taken to avoid contestability becomes merely a way for subjects to complain about their plight. This means contestations of these kinds cannot be handled in a fully automated fashion, because a dialectic exchange is not possible in a meaningful sense between humans and machines. Computational logic can only offer an answer to the "how", whereas a proper response to a contestation must also address the "why" of a given decision (Sarra, 2020). Contestability should include a right to a new decision, compensation of harm inflicted, or reversal (Lyons et al., 2021).

6.1.5 Tools for Scrutiny by Subjects or Third Parties

This feature supports scrutiny by outside actors (decision subjects, indirect stakeholders, third parties) of AI systems, separate from individual decisions. These tools for scrutiny mainly take the form of a range of information resources.

These should contribute to the contestability of the sociotechnical system in its *entirety* (Lyons et al., 2021). The aim is to justify the system as a whole (i.e. "globally"), rather than individual decisions ("locally"). This requires the demonstration of a clear link between high-level objectives (norms external to the technical system) and its implemention. Compliance is established by tracing this link through requirements, specifications, and the code itself.

Documentation should describe the technical composition of the system (Vaccaro et al., 2020). Such documentation may include up-to-date system performance indicators, in particular related to training data and models. Further documentation should describe how the system was constructed (i.e. documentation of the design and development process) (Selbst & Barocas, 2018 in Almada 2019), the role of human decision-makers, group or systemic impacts and how they are safeguarded against (Lyons et al., 2021). Mitchell et al. (2019) and Gebru et al. (2020) offer examples of possible documentation approaches.

Formal proof of compliance may be possible when a system specification can be described unambiguously, and its implementation can be verified (semi-)automatically. However, ML-based systems cannot be described using formal logic. Their performance is better assessed through statistical means (Henin & Le Métayer, 2021).

If a system makes a fully automated decision, it is recommended to include a means of comparing its performance to an equivalent decision-making procedure made by humans (Cowgill & Tucker, 2017 in Almada 2019).

If confidential or sensitive information must be protected that would aid in the assessment of proper system performance, it may be possible to employ "zero-knowledge proofs" in order to provide so-called opaque assurances (Kroll et al., 2016 in Almada 2019).

6.2 Practices

6.2.1 Ex-ante Safeguards

This practice focuses on the earliest stages of the AI system lifecycle, during the business and use-case development phase. It aims to put in place policy-level constraints protecting against potential harms. Developers should make an effort to *anticipate* the impacts of their system in advance (Brkan, 2019; Henin & Le Métayer, 2021; Sarra, 2020), and pay close attention to how the system may "mediate" new and existing social practices (Verbeek 2015 in Hirsch et al., 2017). If after an initial exploration it becomes clear impacts are potentially significant or severe, a more thorough and formalized impact assessment should be performed (e.g. Data Protection Impact Assessments (DPIA)) (Edwards & Veale, 2018; Lyons et al., 2021). Such assessments can also enforce production of extensive technical documentation in service of transparency, and by extension contestability (Bayamlıoğlu, 2021). Any insights from this act of anticipation should feed into the subsequent phases of the AI system lifecycle. Considering AI system development tends to be cyclical and ongoing, anticipation should be revisited with every proposed change (Schot & Rip, 1997 in Kariotis and Mir 2020). If system decisions are found to impact individuals or groups to a significant extent, contestability should be made a requirement (Henin & Le Métayer, 2021). A fairly obvious intervention would be to make contestability part of a system's *acceptance criteria*. This would include the features identified in our framework, first and foremost means of acquiring explanation and human intervention (Almada, 2019; Brkan, 2019; Walmsley, 2021). Questions that must be answered at this point include what can be contested, who can contest, who is accountable, and what type of review is necessary (Lyons et al., 2021).

A final type of ex-ante safeguard is *certification*. This can be applied to the AI system as a software object, by either specifying aspects of its technological design directly, or by requiring certain outputs that enable monitoring and evaluation. It may also be applied to the controlling organization as a whole, which from a sociotechnical perspective is the more desirable option, seeing as how automated decisions cannot be reduced to an AI system's data and model. However, certificates and seals are typically run in a for-profit manner and depend on voluntary participation by organizations. As such they struggle with enforcement. Furthermore, there is little evidence that certificates and seals lead to increased trust on behalf of subjects (Bayamlıoğlu, 2021; Edwards & Veale, 2018).

6.2.2 Agonistic Approaches to ML Development

This practice relates to the early lifecycle phases of an AI system: business and use-case development, design, and training and test data procurement. The aim of this practice is to support ways for stakeholders to "explore and enable alternative ways of datafying and modeling the same event, person or action" (Hildebrandt, 2017 in Almada 2019). An agonistic approach to ML development allows for decision subjects, third parties, and indirect stakeholders to "co-construct the decision-making process" (Vaccaro et al., 2019). The choices of values embedded in systems should be subject to broad debate facilitated by elicitation of the, potentially conflicting, norms at stake (Henin & Le Métayer, 2021). This approach stands in contrast to ex-post mechanisms for contestation, which can only go so far in protecting against harmful automated decisions because they are necessarily reactive in nature (Almada, 2019; Edwards & Veale, 2018). In HCI, a well-established means of involving stakeholders in the development of technological systems is participatory design (Davis, 2009 in Almada 2019). By getting people involved in the early stages of the AI lifecycle, potential issues can be flagged before they manifest themselves through harmful actions (Almada, 2019). Participants should come from those groups directly or indirectly affected by the specific AI systems under consideration. Due to the scale at which many AI systems operate, direct engagement with all stakeholders might be hard or impossible. In such cases, representative sampling techniques should be employed, or collaboration should be sought with third parties representing the interests of stakeholder groups (Almada, 2019). Representation can be very direct (similar to "jury duty"). Or more indirect (volunteer or elected representatives forming a board or focus group) (Vaccaro et al., 2021).

Power differentials may limit the degree to which stakeholders can actually affect development choices. Methods should be used that ensure participants are made aware of and deal with power differentials (Geuens et al., 2018; Johnson, 2003 in Kariotis and Mir 2020).

One-off consultation efforts are unlikely to be sufficient, and run the risk of being reduced to mere "participation theater" or a ticking-the-box exercise. Participation, in the agonistic sense, implies an ongoing adversarial dialogue between developers and decision subjects (Kariotis & Mir, 2020).[2] AI systems, like all designed artifacts, embody particular political values (Winner, 1980 in Crawford 2016). A participatory, agonistic approach should be aimed at laying bare these values, and to create an arena in which design choices supporting one value over an other can be debated and resolved (although such resolutions should always be considered provisional and subject to change) (Kariotis & Mir, 2020). König and Wenzelburger (2021) offer an outline of one possible way of structuring such a process.

[2] For a critique of how participation is not a panacea for all potential harms caused by AI systems, see Sloane et al. (2020).

6.2.3 Quality Assurance During Development

This practice ensures safe system performance during the development phases of the AI system lifecycle. This includes collection of data and training of models, programming, and testing before deployment. A tried and true approach is to ensure the various stakeholder rights, values and interests guide development decisions. Contestability should not be an afterthought, a "patch" added to a system once it has been deployed. Instead developers should ensure the system as a whole will be receptive and responsive to contestations. Care should also be taken to understand the needs and capabilities of human controllers so they will be willing and able to meaningfully intervene when necessary (Kluttz et al., 2018; Kluttz and Mulligan 2019; Leydens & Lucena, 2018 in Almada, 2019; Kariotis & Mir, 2020; Hirsch et al., 2017). Before deploying a system, it can be tested, e.g. for potential bias, by applying the model to datasets with relevant differences (Ploug & Holm, 2020). Given the experimental nature of some AI systems, it may be very challenging to foresee all potential impacts beforehand, on the basis of tests in lab-like settings alone. In such cases, it may be useful to evaluate system performance in the wild using a "living lab" approach (Kariotis & Mir, 2020). In any case, development should be set up in such a way that feedback from stakeholders is collected before actual deployment, and time and resources are available to perform multiple rounds of improvement before proceeding to deployment (Hirsch et al., 2017; Vaccaro et al., 2019, 2020). Developers should seek feedback from stakeholders both with respect to system accuracy, and ethical dimensions (e.g. fairness, justice) (Walmsley, 2021).

6.2.4 Quality Assurance After Deployment

This practice relates to the AI system lifecycle phases following deployment. It is aimed at monitoring performance and creating a feedback loop to enable ongoing improvements. The design concept "procedural regularity" captures the idea that one should be able to determine if a system actually does what it is declared to be doing by its developers. In particular when models cannot be simplified, additional measures are required to demonstrate procedural regularity, including monitoring (Bayamlıoğlu, 2021). System operators should continuously monitor system performance for unfair outcomes both on individuals, and in the aggregate, on communities. To this end, mathematical models can be used to determine if a given model is biased against individuals or groups (Goodman, 2016 in Almada 2019). Monitoring should also be done for potential misuse of the system. Corrections, appeals, and additional contextual information from human controllers and decision subjects can be used as feedback signals for the decision-making process as a whole (Hirsch et al., 2017; Vaccaro et al., 2020). In some cases, feedback loops back to training can be created by means of "reinforcement learning", where contestations are connected to reward functions. In decision-support settings, such signals can also be derived from occurrences where human controllers reject system predictions (Walmsley, 2021).

6.2.5 Risk Mitigation Strategies

This practice relates to all phases of the AI system lifecycle. The aim is to intervene in the broader context in which systems operate, rather than to change aspects of what is commonly considered systems themselves. One strategy is to educate system users on the workings of the systems they operate or are subject to. Such training and education efforts should focus on making sure users understand how systems work, and what their strengths and limitations are. Improving users' understanding of systems may: 1. discourage inappropriate use and encourage adoption of desirable behavior; 2. prevent erroneous interpretation of model predictions; 3. create a shared understanding for the purposes of resolving disputes; and 4. ensure system operators along decision chains are aware of risks and responsibilities (Hirsch et al., 2017; Lyons et al., 2021; Ploug & Holm, 2020; Vaccaro et al., 2019, 2020).

6.2.6 Third-Party Oversight

This practice relates to all phases of the AI system lifecycle. Its purpose is to strengthen the supervising role of trusted third party actors such as government agencies, civil society groups, and NGOs. As automated decision-making happens at an increasingly large scale, it will be necessary to establish new forms of ongoing outside scrutiny (Bayamlıoğlu, 2021; Edwards & Veale, 2018; Elkin-Koren, 2020; Vaccaro et al., 2019). System operators may be obligated to implement model-centric tools for ongoing auditing of systems' overall compliance with rules and regulations (Bayamlıoğlu, 2021). Companies may resist opening up proprietary data and models for fear of losing their competitive edge and users "gaming the system" (Crawford, 2016). Where system operators have a legitimate claim to secrecy, third parties can act as trusted intermediaries to whom sensitive information is disclosed, both for ex-ante inspection of systems overall and post-hoc contestation of individual decisions (Bayamlıoğlu, 2021). Such efforts can be complemented with the use of technological solutions including secure environments which function as depositories for proprietary or sensitive data and models (Edwards & Veale, 2018).

6.3 Contestable AI by Design: Towards a Framework

We have mapped the identified features in relation to the main actors mentioned in the literature (Fig. 2): System developers create *built-in safeguards* to constrain the behavior of AI systems. Human controllers use *interactive controls* to correct or override AI system decisions. Decision subjects use *interactive controls*, *explanations*, *intervention requests*, and *tools for scrutiny* to contest AI system decisions. Third parties also use *tools for scrutiny* and *intervention requests* for oversight and contestation on behalf of individuals and groups.

We have mapped the identified practices to the AI lifecycle phases of the Information Commissioner's Office (ICO)'s auditing framework (Binns & Gallo, 2019) (Fig. 3). These practices are primarily performed by system developers. During

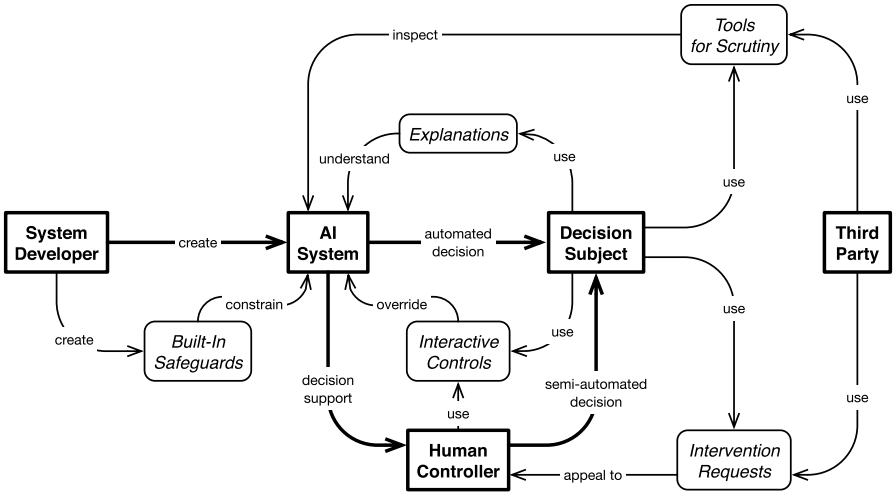

Fig. 2 Features contributing to contestable AI

business and use-case development, *ex-ante safeguards* are put in place to protect against potential harms. During design and procurement of training and test data, *agonistic development approaches* enable stakeholder participation, making room for and leveraging conflict towards continuous improvement. During building and testing, *quality assurance* measures are used to ensure stakeholder interests are centered and progress towards shared goals is tracked. During deployment and monitoring, further *quality assurance* measures ensure system performance is tracked on an ongoing basis, and the feedback loop with future system development is closed. Finally, throughout, *risk mitigation* intervenes in the system context to reduce the odds of failure, and *third party oversight* strengthens the role of external reviewers to enable ongoing outside scrutiny.

7 Discussion

Using a systematic review and qualitative analysis of literature on the design of contestable AI, we have identified five system features and six development practices contributing to AI system contestability. The features are: 1. built-in safeguards against harmful behavior; 2. interactive control over automated decisions; 3. explanations of system behavior; 4. human review and intervention requests; and 5. tools for scrutiny by subjects or third parties. The practices are: 1. ex-ante safeguards; 2. agonistic approaches to ML development; 3. quality assurance during development; 4. quality assurance after deployment; 5. strategies for risk mitigation; and 6. third-party oversight. We used diagrams to capture how features relate to various actors in typical AI systems, and how practices relate to typical AI system lifecycle stages. These features and practices are a step towards more intermediate-level design knowledge for contestable AI. It represents our attempt to take the general principle

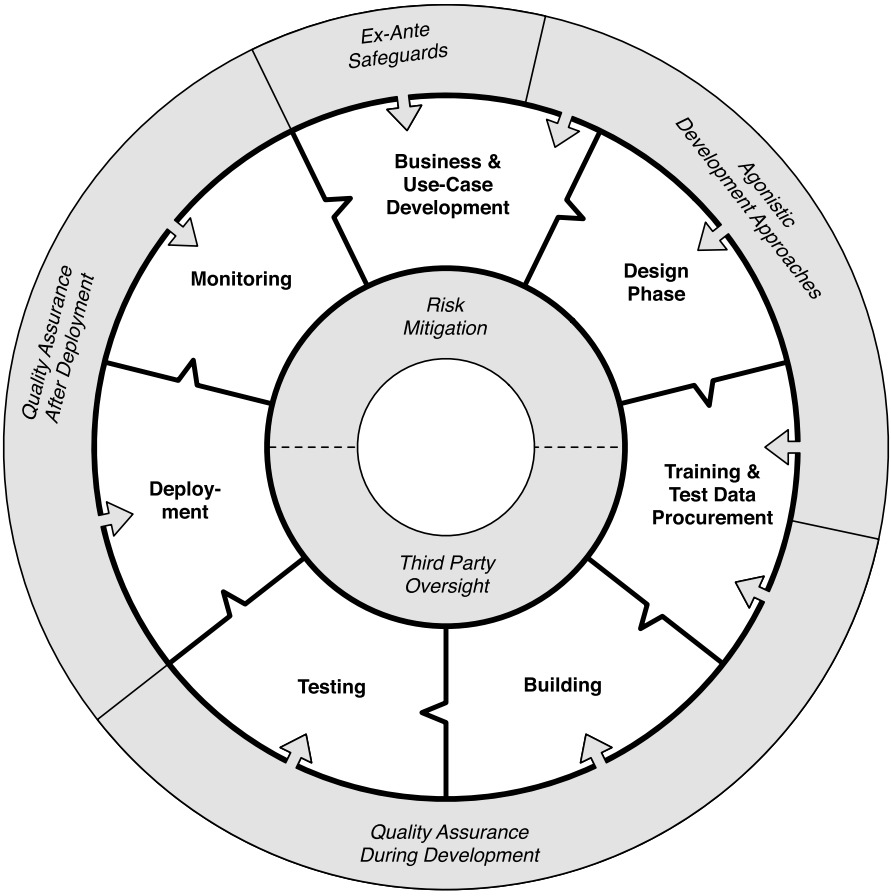

Fig. 3 Practices contributing to contestable AI

of contestability as "open and responsive to dispute" and articulate potential ways in which AI systems, and the practices constituting them, can be changed or amended to support it, with a particular focus on interventions cutting across social and technical dimensions.

Our framework takes a sociotechnical perspective by focusing many of its recommendations on the entangled and volatile nature of AI systems. For example, *interactive control* enables negotiation between artificial and human agents; *explanations* account for the behavior of automated decision-making systems as a whole, not just technical models; *intervention requests* enable a dialectical process between decision subjects and human controllers in close coupling with artificial agents; and *tools for scrutiny* require documentation of not just technical systems but also how they are constructed. Furthermore, *ex ante safeguards* include certification of entire organizations, not just technical systems in isolation; *agonistic design approaches* lay bare how values are embedded in specific

sociotechnical arrangements, creating arenas for stakeholders to co-construct decision-making processes; *QA during development* addresses system volatility through iterative building and testing, possibly in a living lab setting; *QA after deployment* focuses on traceable decision chains across human and artificial agents; and *risk mitigation* educates human controllers and decision subjects on responsible and effective ways of relating to AI system.

The framework has been developed based on a small sample of academic papers. This approach has obvious limitations. There may be gaps caused by lack of coverage in source papers. The papers included approach the subject of contestability from specific fields (e.g. ethics of technology, computer science, law). Many of these papers are not based on empirically validated interventions. While our framework tries to make the translation to practice, most of the papers on which the content of our framework is based are still "context-free". We have developed a framework ready to be tested (and validated) in practice, in specific application contexts. The validation itself was not part of this paper.

Morley et al. (2019) note that many AI ethics tools lack usability in the sense that they are not actionable and do not come with guidance on how they may be put to use in practice. The usability of our own offering here is still limited: We offer diagrams, which are one step up from lists in terms of conceptual richness. The recommendations are on the level of practices and features rather than general principles, making them more actionable. However, we do not offer directions for the use of the framework to actually design contestable AI. Future work should seek to apply the framework in design activities towards the improvement of use situations, or the creation of artifacts embodying the idea of contestable AI for the purpose of further knowledge development.

Many of the themes captured by our framework have also been explored in the literature related to AI accountability. Future efforts may seek to compare our proposed framework to more generic ethical, responsible and accountable AI frameworks (e.g. Cobbe et al., 2021; Hutchinson et al., 2021; Mohseni 2019; Raji et al., 2020).

Our framework assumes no context, or in any case assumes a generic "automated decision-making" setting. It assumes some things are at stake in the decision-making process, typically captured by the phrase "significant impact" on individuals or groups. This covers quite a broad range, but likely does preclude extreme high stakes contexts one finds in e.g. lethal autonomous weapons. Similarly, our framework assumes contexts where time-sensitivity of human intervention is relatively low. That is to say, this framework probably does not cover cases such as shared control of autonomous vehicles. A related research field more focused on these high-stakes and time sensitive scenarios is *meaningful human control* (for which see e.g. Methnani et al., 2021; de Sio & van den Hoven, 2018; Umbrello, 2021; Braun et al., 2021; Verdiesen et al., 2021; Wyatt & Galliott, 2021; Cavalcante Siebert et al., 2022).

Much of our own empirical work is situated in (local) government public services in OECD countries. Some distinctive features of such settings include distribution

of system components across public and private organizations; the duty of care government organizations have towards citizens; and the (at least nominal) democratic control of citizens over public organizations. We expect this framework to hold up quite well in such settings.

A pattern running through all identified features and practices is to avoid attempts to at all cost resolve disputes up front before they arise using some form of compromise or consensus-seeking. Instead, we accept that controversy is at times inevitable, and in fact may even be desirable as a means of spurring continuous improvement. We propose to set up procedural, agonistic mechanisms through which disputes can be identified and resolved. Stakeholders do not need to agree on every decision that goes into the design of a system, or indeed every decision a system makes. However, stakeholders *do* need to agree on procedures by which such disagreements will be resolved. A risk, of course, is that this procedural and adversarial approach is abused to cover for negligence on the part of system designers. This, however, can be addressed by making sure these adversarial procedures include an obligation to account for any decisions leading up to the disagreement under consideration (i.e. ensure decision chains are *traceable*). This adversarial approach should be an effective way to curb the administrative logic of efficiency, and to instead center democratic values of inclusion, plurality, and justice.

8 Concluding Remarks

Subjects of automated decisions have the right to human intervention throughout the AI system lifecycle. Contestable AI by design is an approach that ensures systems respect this right. Most contestable AI knowledge produced thus far lacks adaptability to a design context. Design frameworks are an effective form of knowledge because they are generative and of an intermediate level of abstraction. We analyzed extant literature on contestable AI for system properties enabling contestation. Using visual mapping techniques we synthesized these elements into a design framework. Our framework offers five features and six practices contributing to contestable AI. By thinking in terms of contestability, we close the loop between ex-ante agonistic and participatory forms of anticipation with post-hoc mechanisms for opposition, dissent and debate. In this way, contestability leverages conflict for continuous system improvement.

Appendix 1: Summary of Reviewed Literature

See Tables 2, 3 and 4

Table 2 Included sources and their related features and practices

Source	Features	Practices
Almada (2019)	Built-in safeguards; explanations; intervention requests; tools for scrutiny	Agonistic approaches; ex-ante safeguards; QA after deploy; QA during dev
Aler Tubella et al. (2020)	Explanations; tools for scrutiny	Ex-ante safeguards; QA after deploy
Bayamlıoğlu (2021)	Explanations; interactive control; intervention requests; tools for scrutiny	Ex-ante safeguards; QA after deploy; 3rd party oversight
Brkan (2019)	Explanations; intervention requests	Ex-ante safeguards
Crawford (2016)	Explanations	3rd party oversight
Edwards and Veale (2018)	Explanations; intervention requests	Ex-ante safeguards; 3rd party oversight
Elkin-Koren (2020)	Built-in safeguards; intervention requests	QA during dev; 3rd party oversight
Henin and Le Métayer (2021)	Explanations; intervention requests; tools for scrutiny	Agonistic approaches; ex-ante safeguards
Hirsch et al. (2017)	Explanations; interactive control; intervention requests; tools for scrutiny	Ex-ante safeguards; QA after deploy; QA during dev; risk mitigation
Jewell (2018)	Interactive control	–
Kariotis and Mir (2020)	Tools for scrutiny	Agonistic approaches; ex-ante safeguards; QA during dev
König and Wenzelburger (2021)	Agonistic approaches	–
Lyons et al. (2021)	Explanations; intervention requests; tools for scrutiny	Ex-ante safeguards; risk mitigation; 3rd party oversight;
Ploug and Holm (2020)	Explanations; intervention requests	QA during dev; risk mitigation
Sarra (2020)	Explanations; intervention requests	Ex-ante safeguards
Vaccaro et al. (2019)	Explanations; interactive control; intervention requests; tools for scrutiny	Agonistic approaches; QA during dev; risk mitigation; 3rd party oversight
Vaccaro et al. (2020)	Interactive control; intervention requests; tools for scrutiny	QA after deploy; QA during dev; risk mitigation; 3rd party oversight
Vaccaro et al. (2021)	Explanations; intervention requests	Agonistic approaches; QA after deploy
Walmsley (2021)	Intervention requests	Ex-ante safeguards; QA during dev; QA after deploy

Table 3 Features contributing to contestable AI

Feature	Sources
Built-in safeguards against harmful behavior	Almada (2019), Elkin-Koren (2020)
Interactive control over automated decisions	Bayamlıoğlu (2021), Hirsch et al. (2017), Jewell (2018), Vaccaro et al. (2019, 2020)
Explanations of system behavior	Aler Tubella et al. (2020), Almada (2019), Bayamlıoğlu (2021), Brkan (2019), Crawford (2016), Edwards and Veale (2018), Henin and Le Métayer (2021), Hirsch et al. (2017), Lyons et al. (2021), Ploug and Holm (2020), Sarra (2020), Vaccaro et al. (2019, 2021)
Human review and intervention requests	Almada (2019), Bayamlıoğlu (2021), Brkan (2019), Edwards and Veale (2018), Elkin-Koren (2020), Henin and Le Métayer (2021), Hirsch et al. (2017), Lyons et al. (2021), Ploug and Holm (2020), Sarra (2020), Vaccaro et al. (2019, 2020, 2021), Walmsley (2021)
Tools for scrutiny by subjects or third parties	Aler Tubella et al. (2020), Almada (2019), Bayamlıoğlu (2021), Henin and Le Métayer (2021), Hirsch et al. (2017), Kariotis and Mir (2020), Lyons et al. (2021), Vaccaro et al. (2019, 2020)

Table 4 Practices contributing to contestable AI

Practice	Sources
Ex-ante safeguards	Aler Tubella et al. (2020), Almada (2019), Bayamlıoğlu (2021), Brkan (2019), Edwards and Veale (2018), Henin and Le Métayer (2021), Hirsch et al. (2017), Kariotis and Mir (2020), Lyons et al. (2021), Sarra (2020), Walmsley (2021)
Agonistic approaches to ML development	Almada (2019), Henin and Le Métayer (2021), Kariotis and Mir (2020), König and Wenzelburger (2021), Vaccaro et al. (2019, 2021)
Quality assurance during development	Almada (2019), Elkin-Koren (2020), Hirsch et al. (2017), Kariotis and Mir (2020), Ploug and Holm (2020), Vaccaro et al. (2019, 2020), Walmsley (2021)
Quality assurance after deployment	Aler Tubella et al. (2020), Almada (2019), Bayamlıoğlu (2021), Hirsch et al. (2017), Vaccaro et al. (2020, 2021), Walmsley (2021)
Risk mitigation strategies	Hirsch et al. (2017), Lyons et al. (2021), Ploug and Holm (2020), Vaccaro et al. (2019, 2020)
Third-party oversight	Bayamlıoğlu (2021), Crawford (2016), Edwards and Veale (2018), Elkin-Koren (2020), Lyons et al. (2021), Vaccaro et al. (2019, 2020)

Acknowledgements The authors would like to thank the reviewers for their constructive comments.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by KA. The first draft of the manuscript was written by KA and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research was supported by a grant from the Dutch National Research Council NWO (grant no. CISC.CC.018).

Data Availability Materials and data are available at the 4TU.ResearchData repository under DOI 10.4121/15350118.

Code availability Not applicable.

Declarations

Conflict of interest Not applicable.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

Aler Tubella, A., Theodorou, A., Dignum, V., et al. (2020). Contestable black boxes. In V. Gutiérrez-Basulto, T. Kliegr, A. Soylu, et al. (Eds.), *Rules and reasoning* (Vol. 12173). Springer.

Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law*, ICAIL 2019, pp 2–11

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society, 20*(3), 973–989.

Applebee, A. N., & Langer, J. A. (1983). Instructional scaffolding: Reading and writing as natural language activities. *Language Arts, 60*(2), 168–175 http://www.jstor.org/stable/41961447.

Bayamlıoğlu, E. (2021). The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation". Regulation and Governance.

Binns, R., & Gallo, V. (2019). An overview of the Auditing Framework for Artificial Intelligence and its core components. https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

Braun, M., Bleher, H., & Hummel, P. (2021). A leap of faith: Is there a formula for "trustworthy'' AI? *Hastings Center Report, 51*(3), 17–22. https://doi.org/10.1002/hast.1207.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Brkan, M. (2019). Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International Journal of Law and Information Technology, 27*(2), 91–121.

Cavalcante Siebert, L., Lupetti, M. L., & Aizenberg, E., et al. (2022). Meaningful human control: Actionable properties for AI system development. AI and Ethics.

Chiusi, F., Fischer, S., & Kayser-Bril, N., et al. (2020). Automating Society Report 2020. Tech. rep., Algorithm Watch, https://automatingsociety.algorithmwatch.org

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Virtual Event, pp 598–609, https://doi.org/10.1145/3442188.3445921

Cowgill, B., & Tucker, C. (2017). Algorithmic bias: A counterfactual perspective. Working Paper: NSFTrustworthy Algorithms p 3. http://trustworthy-algorithms.org/whitepapers/Bo Cowgill.pdf

Crawford, K. (2016). Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values, 41*(1), 77–92

Crawford, K., Dobbe, R., & Dryer, T., et al. (2019). AI now 2019 report. Technical report, AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.html

Davis, J. (2009). Design methods for ethical persuasive computing. In *Proceedings of the 4th international conference on persuasive technology*. Association for Computing Machinery, Persuasive '09.

de Sio, F. S., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers Robotics AI, 5,* 1–14.

Dorst, K., & Cross, N. (2001). Creativity in the design process: Co-evolution of problem-solution. *Design Studies, 22*(5), 425–437.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing, 8*(1), 19–30.

Edwards, L., & Veale, M. (2018). Enslaving the algorithm: From a "right to an explanation'' to a "right to better decisions''? *IEEE Security & Privacy, 16*(3), 46–54.

Elkin-Koren, N. (2020). Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society, 7*(2), 205395172093,229.

Franssen, M. (2015). Design for values and operator roles in sociotechnical systemsociotechnical systems. In: van den Hoven J, Vermaas PE, van de Poel I (Eds.) Handbook of Ethics, Values, and Technological Design. Springer, pp 117–149, https://doi.org/10.1007/978-94-007-6970-0_8

Gebru, T., Morgenstern, J., & Vecchione, B., et al. (2020). Datasheets for datasets. arXiv:1803.09010 [cs]

Geuens, J., Geurts, L., Swinnen, T. W., et al. (2018). Turning tables: A structured focus group method to remediate unequal power during participatory design in health care. In *Proceedings of the 15th participatory design conference: Short papers, situated actions, workshops and tutorial - Volume 2*. ACM, Hasselt and Genk, pp 1–5.

Goodman, B. (2016). Economic models of (algorithmic) discrimination. In *29th conference on neural information processing systems*

Henin, C., & Le Métayer, D. (2021). Beyond explainability: Justifiability and contestability of algorithmic decision systems. AI & Society

Hildebrandt, M. (2017). Privacy as protection of the incomputable self: Agonistic machine learning. *SSRN Electronic Journal* 1–33.

Hirsch, T., Merced, K., Narayanan, S., et al. (2017). Designing contestability: Interaction design, machine learning, and mental health. In *DIS 2017 - Proceedings of the 2017 ACM conference on designing interactive systems*. ACM Press, pp 95–99.

Höök, K., Karlgren, J., & Wærn, A., et al. (1998). A glass box approach to adaptive hypermedia. In: Brusilovsky P, Kobsa A, Vassileva J (Eds.) Adaptive hypertext and hypermedia. Springer, pp 143–170, https://doi.org/10.1007/978-94-017-0617-9_6

Höök, K., & Löwgren, J. (2012). Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Transactions on Computer-Human Interaction, 19*(3), 1–18.

Hutchinson, B., Smart, A., & Hanna, A., et al. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. ACM, Virtual Event Canada, pp 560–575.

Jewell, M. (2018). Contesting the decision: Living in (and living with) the smart city. *International Review of Law, Computers and Technology*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Johnson, D. W. (2003). Social interdependence: Interrelationships among theory, research, and practice. *American Psychologist, 58*(11), 934–945.

Kamarinou, D., Millard, C., & Singh, J. (2016). Machine learning with personal data. *Queen Mary School of Law Legal Studies Research Paper, 1*(247), 23.

Kariotis, T., & Mir, D. J. (2020). Fighting back algocracy: The need for new participatory approaches to technology assessment. In *Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 2*. ACM, Manizales Colombia, pp 148–153.

Katell, M., Young, M., & Dailey, D., et al. (2020). Toward situated interventions for algorithmic equity: Lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery*, pp 45–55, https://doi.org/10.1145/3351095.3372874

Kluttz, D., Kohli, N., & Mulligan, D. K. (2018). Contestability and professionals: From explanations to engagement with algorithmic systems. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3311894.

Kluttz, D. N., & Mulligan, D. K. (2019). Automated decision support technologies and the legal profession. *Berkeley Technology Law Journal, 34*(3), 853. https://doi.org/10.15779/Z38154DP7K.

Kluttz, D. N., Mulligan, D. K., Mulligan, D. K., et al. (2019). Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3311894.

König, P. D., & Wenzelburger, G. (2021). The legitimacy gap of algorithmic decision-making in the public sector: Why it arises and how to address it. *Technology in Society, 67*(101), 688.

Kroes, P., Franssen, M., van de Poel, I., et al. (2006). Treating socio-technical systems as engineering systems: Some conceptual problems. *Systems Research and Behavioral Science, 23*(6), 803–814.

Kroll, J. A., Barocas, S., Felten, E. W., et al. (2016). *Accountable algorithms. U Pa L Rev, 165,* 633.

Leahu, L. (2016). Ontological surprises: A relational perspective on machine learning. In *Proceedings of the 2016 ACM conference on designing interactive systems*. ACM, pp 182–186

Leydens, J. A., & Lucena, J. C. (2018). *Engineering justice: Transforming engineering education and practice*. IEEE PCS Professional Engineering Communication Series. Wiley.

Löwgren, J., Gaver, B., & Bowers, J. (2013). Annotated Portfolios and other forms of intermediate- level knowledge. Interactions pp 30–34.

Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–25.

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*

Matias, J. N., Johnson, A., Boesel, W. E., et al. (2015). Reporting, reviewing, and responding to harassment on twitter. https://doi.org/10.48550/ARXIV.1505.03359

Mendoza, I., & Bygrave, L. A. (2017). The right not to be subject to automated decisions based on profiling. In *EU Internet Law*. Springer, pp 77–98

Methnani, L., Aler Tubella, A., & Dignum, V., et al. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence* 4.

Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. ACM, pp 220–229

Moher, D., Liberati, A., Tetzlaff, J., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000,097.

Mohseni, S. (2019). Toward design and evaluation framework for interpretable machine learning systems. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*. ACM, pp. 553–554.

Morley, J., Floridi, L., Kinsey, L., et al. (2019). From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26,* 2141–2168.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society, 20*(11), 4366–4383.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus, 140*(4), 32–48.

Norman, D. A., & Stappers, P. J. (2015). DesignX: Complex sociotechnical systems. *She Ji: The Journal of Design, Economics, and Innovation, 1*(2), 83–106.

Novick, D. G., & Sutton, S. (1997). What is mixed-initiative interaction. In *Proceedings of the AAAI spring symposium on computational models for mixed initiative interaction*, p 12.

Obrenović, Ž. (2011). Design-based research: What we learn when we engage in design of interactive systems. *Interactions, 18*(5), 56–59.

Ouzzani, M., Hammady, H., Fedorowicz, Z., et al. (2016). Rayyan: A web and mobile app for systematic reviews. *Systematic Reviews, 5*(1), 210.

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics: A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine, 107*(101), 901.

Raji, I. D., Smart, A., White, R. N., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. FAT* 2020—Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency pp 33–44

Ribeiro, M. T., Singh, S., & Guestrin, C. (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*. ACM, pp 1135–1144.

Rouvroy, A. (2012). The end(s) of critique: Data-behaviourism vs. due-process. In: Hildebrandt M, De Vries E (Eds.) *Privacy, due process and the computational turn. Philosophers of Law Meet Philosophers of Technology*.

Salehi, N., Teevan, J., Iqbal, S., et al. (2017). Communicating context to the crowd for complex writing tasks. In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. ACM, Portland, pp 1890–1901.

Sandvig, C., Hamilton, K., Karahalios, K., et al. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. In: *Data and discrimination: Converting critical concerns into productive inquiry*.

Sarra, C. (2020). Put dialectics into the machine: Protection against automatic-decision-making through a deeper understanding of contestability by design. *Global Jurist, 20*(3), 20200,003.

Schot, J., & Rip, A. (1997). The past and future of constructive technology assessment. *Technological Forecasting and Social Change, 54*(2–3), 251–268.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *SSRN Electronic Journal*.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems, 10*(4), 1–31.

Sloane, M., Moss, E., Awomolo, O., et al. (2020). Participation is not a design fix for machine learning. arXiv:2007.02423 [cs]

Stolterman, E., & Wiberg, M. (2010). Concept-driven interaction design research. *Human-Computer Interaction, 25*(2), 95–118.

Suchman, L. (2018). Corporate accountability. https://robotfutures.wordpress.com/2018/06/10/corporate-accountability/

Tickle, A., Andrews, R., Golea, M., et al. (1998). The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks, 9*(6), 1057–1068.

Tonkinwise, C. (2016). The interaction design public intellectual. *Interactions, 23*(3), 24–25).

Umbrello, S. (2021). Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: A two-tiered approach. *Ethics and Information Technology, 23*(3), 455–464.

Vaccaro, K., Karahalios, K., Mulligan, D. K., et al. (2019). Contestability in algorithmic systems. In *Conference companion publication of the 2019 on computer supported cooperative work and social computing*. ACM, pp 523–527

Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). At the end of the day Facebook does what it wants: How users experience contesting algorithmic content moderation. In *Proceedings of the ACM on human-computer interaction* 4.

Vaccaro, K., Xiao, Z., Hamilton, K., et al. (2021). Contestability for content moderation. In: Proceedings of the ACM on human-computer interaction, pp 1–28.

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. Minds and machines

Verbeek, P. P. (2015). Beyond interaction: A short introduction to mediation theory. *Interactions, 22*(3), 26–31.

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. *Minds and Machines, 31*(1), 137–163.

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & SOCIETY, 36*(2), 585–595.

Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136.

Wyatt, A., & Galliott, J. (2021). An empirical examination of the impact of cross-cultural perspectives on value sensitive design for autonomous systems. *Information, 12*(12), 527.