Contestable Artificial Intelligence

Constructive design research for public artificial intelligence systems that are open and responsive to dispute

Alfrink, Kars

DOI
10.4233/uuid:0b014ee6-67a8-4c59-8598-4bfd771595a3

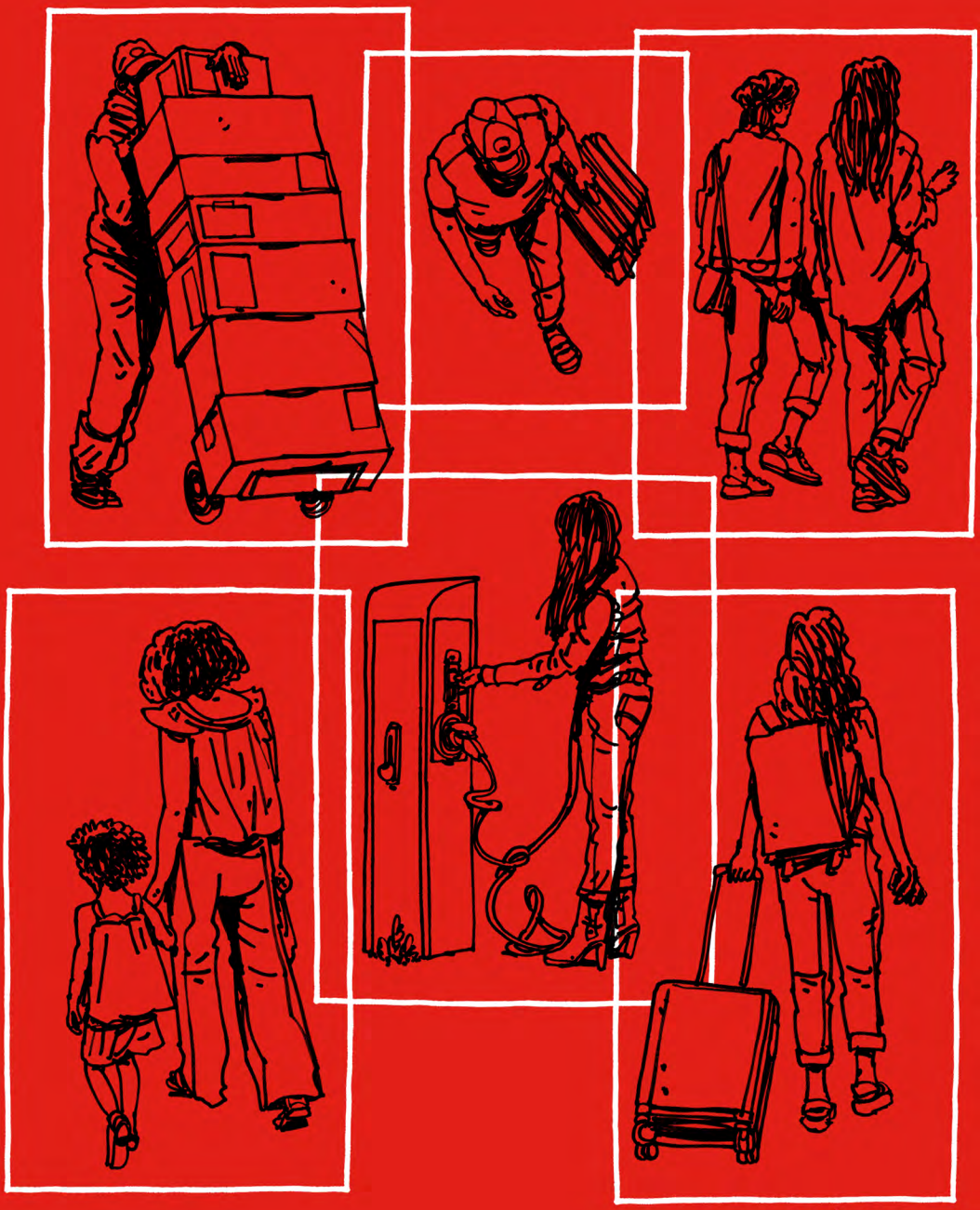
Publication date
2024

Document Version
Final published version

Citation (APA)
Alfrink, K. (2024). *Contestable Artificial Intelligence: Constructive design research for public artificial intelligence systems that are open and responsive to dispute*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:0b014ee6-67a8-4c59-8598-4bfd771595a3

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

CONTESTABLE ARTIFICIAL INTELLIGENCE



KARS ALFRINK

CONSTRUCTIVE DESIGN RESEARCH FOR PUBLIC ARTIFICIAL INTELLIGENCE SYSTEMS THAT ARE OPEN AND RESPONSIVE TO DISPUTE.

Contestable Artificial Intelligence

Constructive design research for public artificial intelligence systems that are open and responsive to dispute.

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Thursday, 23 May 2024 at 15:00 o'clock

by

Christiaan Pieter ALFRINK
European Media Master of Arts in Gaming,
Utrecht School of the Arts, The Netherlands,
validated by the University of Portsmouth, United Kingdom
born in Brederwiede, The Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof.dr. G.W. Kortuem	Delft University of Technology, promotor
Prof.dr.mr.ir. N. Doorn	Delft University of Technology, promotor
Dr.ir. A.I. Keller	Delft University of Technology, copromotor

Independent members:

Prof.dr. E.A. van Zoonen	Erasmus University Rotterdam
Prof.dr. J. Löwgren	Linköping University
Prof.dr. M.V. Dignum	Umeå University
Prof.dr.ir. I.R. van de Poel	Delft University of Technology
Prof.dr. P.J. Stappers	Delft University of Technology, reserve member

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

Printed by:	De Kijm & Zonen (dekijm.nl)
ISBN:	978-94-6366-833-0
DOI:	10/mgtv
Cover & interior art:	Joost Stokhof (jooststokhof.nl)
Layout:	Kars Alfrink
Typeface:	IBM Plex (ibm.com/plex)

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Governments increasingly use artificial intelligence (AI) to support or entirely automate public service decision-making. As the use of AI in public sector decision-making increases, so do concerns over its harmful social consequences, including the undermining of the democratic rule of law and the infringement of fundamental human rights to dignity and self-determination [e.g. 61, 67]. Increasing systems' *contestability*—which I define as openness and responsivity to dispute—is a way to counteract such harms. Contestable AI is a small but growing field of research [9, 55, 146, 153, 298, 341]. However, thus far, much of the research has focused on general principles rather than application in practice.

This thesis' central aim is to explore what sociotechnical design interventions increase the contestability of public AI systems. In this introductory chapter, I first motivate the research by investigating how public AI systems specifically impact human autonomy. I introduce the example cases that play a central role in the subsequent chapters and make the case for contestability as a system quality that supports autonomy. Subsequently, I review the literature on transparency and explainability, contestability, public AI, and agonistic pluralism. I highlight the thesis' knowledge gaps and research questions throughout the review. Next, I lay out the research approach, which I frame as constructive design research, and describe the design and analysis methods. I close with an overview of the remaining chapters.

1.1 Motivation

In this section, I motivate the research of this thesis. I begin by briefly describing the research context, namely governments' use of algorithmic systems to automate or support the execution of policy, which I call *public AI*. I then discuss how these systems impact an essential human value, namely *autonomy*. Next, I introduce the *three cases* that were the basis of the research in this thesis: (1) smart

electric vehicle (EV) charging, (2) camera cars, and (3) fraud risk-scoring models. I use these cases to illustrate how public AI can harm human autonomy. Finally, I make the case for *contestability*, a system quality that protects people's autonomy. The identification of contestability as a desirable quality then sets the scene for the main research aim of this thesis: *To explore what sociotechnical system properties increase the contestability of public AI systems.*

1.1.1 AI for Public Administration and Urban Governance

I situate this work in the context of public AI, which I define, following Suchman [326] and Nouws et al. [265] as *the application of adaptive data analysis and processing to enhance, assist, or automate decision-making in the public sector*.[1]

Governments increasingly make use of public AI [263, 315, 329, 361, 372]. Application areas include child protection, public housing, health, social protection, security, and taxation [48, 86, 230]. Main concerns include transparency [48, 86, 106], data collection politics [230, 277], and impact on public sector work [109, 300, 301, 346].

A related field is *urban AI* [68, 220, 221], which delves into AI's role in the built environment. Application areas here are mainly related to mobility solutions such as electric vehicle charging, autonomous vehicles, and parking systems [6, 219, 299]. This research examines AI's influence on urban experiences, intertwining AI ethics with urban design ethics [220]. The focus on *spatial justice* [142, 206, 307, 314] is more pronounced in *urban* AI studies, complementing the procedural and distributive justice discussions that are more prevalent in public AI research [30, 204]. Procedural and distributive justice are concerned

1. *A note on terminology:* In this thesis, I use "artificial intelligence" and "AI" as an umbrella term to refer to various practices of adaptive data analysis and processing for human-machine decision-making. This use has become quite common in lay discourse. While suffering from inflationary hype, it has also become a helpful boundary object term that links conversations in diverging academic fields. In some cases, when I discuss AI, this can involve machine learning (ML), but this is not required. Where relevant, I distinguish between such stochastic (probabilistic) approaches and their deterministic (rule-based) counterparts. Two other common terms are "algorithms" and "automated decision-making" (sometimes "algorithmic decision-making"). The use of "algorithm"—which strictly refers to any finite set of instructions for solving a mathematical problem—to refer to complex sociotechnical systems that involve human and machine actors is a bit of a misnomer but quite common in lay discourse. I generally avoid its use in this thesis. I use "algorithmic system" to refer to any sociotechnical system that uses computation as part of its decision-making. "Automated–" and "algorithmic decision-making" are common terms in the literature and highlight the particular purpose actors use AI technology for, which is also the focus of this thesis: decision-making in the public sector. Some take issue with the "automated" designation since, more often than not, the systems referred to involve a blend of human and machine agents. For this reason, I usually restrict its use in these writings to refer to fully automated decision-making *only* and use "algorithmic" to refer to all other cases.

with the perceived fairness of decision-making processes and the distribution of resources, respectively. Spatial justice is concerned with how space is used and how decisions about the use and design of particular spaces are determined.

1.1.2 Autonomy

AI systems can potentially erode individual autonomy, a key concern highlighted in numerous AI ethics guidelines. However, the definition of autonomy and its potential compromise by AI is often vague. A clear understanding is crucial as varying interpretations of autonomy lead to different recommendations. Fjeld et al. [108] points out that autonomy typically underpins the concept of human control over technology.

I define *autonomy* following Prunkl [283] as the ability to self-govern effectively, which involves *authenticity*—holding beliefs and values free from external influence—and *agency*—having the capability to act based on one's beliefs and values. These two elements dictate how autonomy is protected or promoted. Christman [62] suggests that a decision or desire is only authentic if one does not feel estranged from it after thoughtful consideration.

Rubel et al. [292] delve into how algorithmic systems impact autonomy. They argue that these systems must be ones individuals can *reasonably endorse*, meaning they align with personal goals or adhere to fair terms of social cooperation. This endorsement depends on the system's reliability, the subject's responsibility for its inputs, the stakes involved, and the distribution of its impact across different groups. I will apply this reasonable endorsement test to examine the effects of public AI systems on autonomy in the following section.

To respect autonomy, individuals should have access to information that supports their practical ability to carry out their plans and cognitive ability to evaluate their circumstances [292]. From this, we can infer principles for informed decision-making and information control. In the next section, I will use these principles to explore necessary design changes in public AI systems to respect autonomy.

Further, Rubel et al. [292] look at the necessary conditions for autonomy, which include freedom—understood as "ecological non-domination." This conception combines aspects of negative, positive, and republican freedom. To be genuinely free, individuals must be able to self-govern, have quality agency, and be free from others' domination. Challenges to freedom include emotional influence, cognitive limitations, and social impacts—all of which others can manipulate.

Finally, Rubel et al. [292] addresses the responsibilities that accompany being autonomous. They note how those deploying AI can use it to conceal accountability—a practice termed "agency laundering," the act of attributing causal responsibility to an actor other than oneself. They also emphasize citizens' role in legitimating political actions, asserting that for a government policy to be legitimate, it must be democratically willed (within normative bounds to avoid arbitrariness), *or* rely on "normative authority" [275] *and* meet the "access constraint," which ensures citizens can form beliefs about policies with a sufficient degree of agency. When algorithmic systems curtail autonomy, they also impede this process of legitimation.

1.1.3 Example Cases: Smart EV Charging, Camera Cars, and Risk Models

In this thesis, the empirical work is grounded in three cases, all of which occur in the city of Amsterdam, the Netherlands: (1) a smart EV charging system, (2) camera cars used for vehicular urban sensing, and (3) a risk scoring model used for enforcement of illegal vacation rentals.

I will briefly describe each, in turn, to add context to the notion of public AI. I will also highlight how autonomy is impacted for each case, using Rubel et al.'s reasonable endorsement test. Remember that subjects' ability to endorse a system depends on its reliability, the degree to which it uses inputs that they can be held responsible for, the stakes involved, and the distribution of burdens across groups.

I will also use principles for informed practical and cognitive agency to sketch out potential interventions that will make these example public AI systems more respectful of people's autonomy by providing particular information and means of control.

Smart EV Charging

Amsterdam operates 2503 charging stations with 4974 charging points (Figure 1.1). Four hundred fifty-two stations (904 points) are part of a smart charging system called Flexpower, which increases charge speed when solar energy is available. In 2016, a design study was commissioned to make smart charging transparent for EV drivers. This study led to the Transparent Charging Station prototype, which uses priority schemes to give shared EVs priority to charge faster. A follow-up project, UI for Smart EV Charging, developed a transparency interface for existing Flexpower charge points and was aimed at studying the feasibility, usefulness, usability, and desirability of transparency.

Figure 1.1
Electric cars at a public charging station in Amsterdam. (Photo: Michiel Wijnbergh)

EV drivers may reject the system because it bases charge speed on factors they cannot be held responsible for—grid capacity, availability of renewable energy—and which they feel it unreasonable to be impacted by. Drivers may still consider this use of factors outside their control acceptable because the system is reliable, and the stakes are relatively low—curbside charging generally is a secondary means of EV charging. Another reason drivers may reject the system is if burdens are distributed unevenly across groups. This distribution could result from variable adaptation for grid capacity between city areas. However, the system currently models capacity at the level of the entire city.

This case directly addresses informational provisions in support of agency—the transparency interface aimed to explain the reasons for the charge speed provided in real-time. As will be shown in Chapter 2, shortcomings in the eyes of EV drivers of this information can be understood in part as a mismatch between the information provided and drivers' practical agency needs. That is to say, drivers were insufficiently supported in planning and acting in line with their values. This mismatch was partly due to the timing and modality

of the information provided and, in part, to a lack of control over charging station behavior.

Camera Cars

Amsterdam and 12 other Dutch municipalities have started using camera cars for parking monitoring and enforcement (Figure 1.2). The system checks if parked cars have paid their parking fee or have a permit. It captures images of license plates and uses computer vision algorithms to recognize them. Payment must be made within 5 minutes, or a parking inspector will review the situation based on four photos. A parking fine is issued if no exceptional circumstances apply. Amsterdam also uses camera cars to detect stolen vehicles and those with a claim from the police or public prosecutor.

Figure 1.2
A parking monitoring camera car driving through the streets of Amsterdam. (Photo: Robin Utrecht)

Sticking with parking enforcement camera cars, drivers may reject the system based on its (lack of) reliability, the high stakes involved, and the relative burden imposed on some groups. Misread license plates may cause unwarranted

fining or failure to detect exceptional circumstances (e.g., curbside unloading). Fines can be significant, and a failure to satisfy them on time can lead to notable increases. System unreliability can impact certain groups more than others. For example, those who rely on curbside unloading for their daily routines—parents who drop off children at daycares or schools—or their livelihoods—parcel deliverers—may have to deal more frequently with unwarranted fines.

Given the system's unreliability, what is essential here is informational control—drivers subjected to erroneous fines should be provided with the information required to determine the reasons for the fine and allowed to make corrections to the data on which this decision was based. The history of the Amsterdam parking enforcement camera car illustrates this issue. Only after initial deployment did it become clear that circumstances might warrant exemption. This realization led to a change in procedure, by which human review was added as a final step to positive system detections. Furthermore, a custom web interface with some integration with the primary algorithmic system was added to enable drivers to review the images from the camera car that led to the fine and to object if they felt the fine was unwarranted.

Fraud Risk Scoring Models

Amsterdam is struggling with mass tourism (Figure 1.3). Visitor levels have rapidly recovered to pre-pandemic levels, and the practice of illegal vacation rental properties makes it difficult to control visitor flows. To address this issue, the city introduced a pilot system in 2020 that helps screen reports of possible illegal vacation rentals. The system calculates the probability of housing fraud using a model created using random forest regression and historical data on investigated reports. A civil servant decides whether or not to investigate based on the report, risk score, and explanation, and enforcement officers conduct the investigation. However, high fines have led to concerns about disproportionate enforcement for minor violations.

Those subjected to this system by being reported on and subsequently selected for investigation could object to this system mainly based on data being included that they cannot be held responsible for, such as gender, birth date, family composition, and the property's address and characteristics. As indicated, there are significant stakes involved—being visited by inspectors can be an unpleasant experience in and of itself. Fines for violations are significant (ranging from €8,700 to €21,750). These high stakes compound the issue of responsibility.

Figure 1.3
Tourists walk through the center of Amsterdam with suitcases. Someone holding a phone with the Airbnb app in the foreground. (Photo: Robin Utrecht)

In particular, cognitive agency may be at stake with this system. If subjects are not provided with information on the fact that they have been reported on by someone else and an algorithmic system has scored the report as high-risk, being investigated can feel like coming entirely out of the blue. Subjects will struggle to exercise sufficient evaluative control over their lives without such information.

1.1.4 From Autonomy, via Human Control, to Contestability

This brief discussion of this thesis' example cases has illustrated a variety of public AI systems currently in use and how they might impact subjects' autonomy.

Ethical and rights-based approaches to AI typically address autonomy with principles such as transparency, explainability, accountability, and, most notably, for our purposes, human control of technology. Subjects should be provided with explanations of AI decisions and notified when a system decides about them or when interacting with an AI system. Subjects should be given

the ability to appeal AI decisions. Finally, subjects should be entitled to human review of AI decisions or to opt out of automated decisions entirely. In general, AI should be developed and implemented to allow subjects to intervene in system actions [108].

The system quality that affords such human control I choose to call *contestability*. For this, we initially took inspiration from Hirsch et al. [153], who were among the first to articulate contestability concerning the HCI design of AI systems.[2] They use a case study of an automated assessment and training tool for psychotherapists. The system, called CORE-MI, evaluates counseling sessions. The counselors involved expressed a desire to be able to contest the system reports. Hirsch et al. claim contestability is particularly important for systems that evaluate human performance. They frame contestation as humans "challenging machine predictions." Hirsch et al. argue that we should recognize that our models are and will continue to be fallible and that the risks of failure can be high. They discuss how it is necessary to take responsibility for how technologies mediate human-world and human-human relations [347]. To address such issues, Hirsch et al. argue for a commitment to improving system accuracy; ensuring outputs are explainable, traceable, and opposable; training users to understand the limitations of AI tools better; monitoring for bias and misuse; and enabling users to ask questions and lodge complaints.

What political ideal could underwrite a project of contestable AI? Framing human control of technology in the interest of autonomy as contestability is in keeping with *agonistic pluralism* [249–255]. Agonists argue for a "return of the political" [255] and for perceiving competition as something to be celebrated. Agonistic pluralism is characterized by a commitment to radical pluralism (a diversity of values is constructive, not needing resolution), a tragic view of the world (where conflict is ineradicable and intrinsic to social relations), and a conviction that conflict can be productive [216].

Finally, the acknowledgment of fallibility and emphasis on provisionality argued for by Hirsch et al., and echoed by the agonists also resonates with the work of Collingridge [64] who in the context of technology policy and "the social control of technology" forcefully argued against the focus on anticipating errors up-front to the exclusion of other measures. Given that decisions about social technologies are performed under "ignorance," we should continuously monitor systems for errors and ensure our ability to revise systems and for decisions to

2. Here I write "we" because this inspiration was the product of conversations between my promotors and myself.

be reversed whenever possible.

To summarize, public AI systems can harm people's autonomy. Principles seeking to address these include transparency and explainability, accountability, and *human control of technology*. Filtered through an agonistic view, one that faces up to our fallibility and taking responsibility for technology mediations, I propose *contestability* as a system quality that ensures public AI systems respect people's autonomy.

The argument for contestability leads us to a formulation of the overall research question of this thesis: *What socio-technical design interventions enhance the contestability of public AI systems?*

1.2 Literature Review, Knowledge Gaps, and Research Questions

This section briefly reviews selected literature on transparency, contestability, public AI, and agonistic pluralism and identifies a series of gaps and accompanying research questions. This narrative tracks the thesis' structure and the chronological order in which this research unfolded.

1.2.1 Transparency

I review transparency first from prominent philosophical accounts and then look at empirical work on the topic in HCI specifically. I close with the knowledge gap and accompanying research question.

Transparency's Role in Accountability and Legitimacy

Transparency is frequently seen as a key way of ensuring AI systems' accountability and legitimacy [28, 51, 70, 106]. Sources of AI system opacity include corporate secrecy, subjects' illiteracy, and, most challengingly, the divergence between how humans understand the world and the representations that machines build in the form of ML models [51].

Transparency in the form of *explanations* for subjects of automated decision-making emerged with the implementation of the European Union General Data Protection Regulation (GDPR) in 2018. However, the so-called "right to explanation" contained in the GDPR is limited by its constrained scope and ambiguous language [352].

One aim of transparency is to ensure the *accountability* of organizations that deploy AI. Accountability can be conceptualized as the obligation on the

part of system operators to provide subjects with a justification for their conduct, where the operator faces sanction if the subject deems an account inadequate [28]. Full transparency in the interest of accountability is limited by potential harms, including subjects' loss of privacy, risk of strategic behavior, loss of companies' competitive advantage, and inherent opacity of specific technical approaches [70].

Closely related to accountability is the aim of *legitimacy*. Suchman [328] defines legitimacy as "a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions."[3,4]

A way to enable subjects to legitimate AI systems without suffering the harms associated with full transparency [70] is to provide *justifications* instead.[5] Justifications are a limited form of explanations that offer an account of "what the decision is, on which grounds it has been made, and in doing so, identify who the responsible actor is" [106].

Transparency's Effects on Understanding, Trust, and Control

In HCI, transparent and explainable AI has become a prominent object of study [12, 27, 96, 188, 284, 311].

To make sense of the literature on transparent, explainable, and interpretable AI in HCI, we can distinguish between at least three potential audiences for explanations: (1) *Developers.* Those who build models and the systems that embed them, (2) *Users.* Those who use model predictions as part of their decision-making. (3) *Subjects.* Those impacted by decisions determined either wholly or partially by models. Predictably, much of HCI work focuses on the experiences of model *users*. Developers and subjects figure less prominently. This thesis is concerned chiefly with the fate of subjects.

An early definition of transparency, in the context of recommender systems, is "user understanding of why a particular recommendation was made" [311]. This focus on *understanding* has been enduring in much of HCI research. Users like and feel more confident about recommendations they perceive as transparent [311]. Explanation contents and form influence recipient understanding. One study finds that the best-performing explanations are those that expose model internals and allow for interactive exploration of model behavior.

3. I follow Henin and Le Métayer [146] in using this definition who, following Waldman [353], consider it general enough to apply to AI.
4. This definition is consistent with the account of legitimation provided by Rubel et al. [292].
5. This account is consistent with the conceptualization of accountability provided by Binns [28].

Interactive explanations do, however, require a higher time investment than static ones [60].

A particular approach to transparency is through *interpretable models*, which, in contrast to black box models that can only be explained post-hoc, are human-understandable ex-ante (e.g., because they are rules-based) [293, 306, 368].

Evaluating the degree to which transparency influences the inclination of users to follow model predictions and their ability to detect mistakes, Poursabzi-Sangdeh et al. [282] find that understanding is best served by models that use fewer features and are highly interpretable. However, error detection is negatively affected by models proactively accompanied by explanations, likely due to information overload. Better interpretability does *not* lead to higher user compliance [282].

Transparency is often pursued to increase awareness and *trust* [188]. The relationship between the amount of information provided and subjects' trust is not straightforward. Too much information can erode trust. Transparency measures really only make a difference when expectations are violated [188]. Trust does not appear to increase regardless of the type of explanation used or level of understanding [60]. Data-centric explanations, describing training data to end users, positively impact trustworthiness assessment and aid in assessing fairness [12].

Explanations' effectiveness in increasing user control over system outputs is limited, suggesting that transparency alone does not enable users to influence or change system operations significantly [284]. For example, Airbnb hosts face anxiety due to the uncertainty caused by a partially transparent algorithm, pointing to the need for a design that offers both information and control without conflicting with the platform's objectives [173]. This case exemplifies the broader tension between transparency and company interests, such as the prevention of strategic behavior and protection of intellectual property [70, 352]. Similarly, increased transparency in the algorithm used by Yelp—a crowd-sourced business review platform—led users to either manipulate their behavior to please the algorithm or to disengage from the platform altogether, reflecting a lack of 'voice' or 'loyalty' as described in Hirschman's model [154].

The Sociotechnical, Contextual, and Relational View of Transparency

The transparency of models in isolation is limited in various ways, which should temper our optimism. Making models transparent is not the same as holding to account the entire "sociotechnical assemblage" they comprise. These can never be wholly seen into, held still, or fully traced. Rather than seeing inside,

an ability to "see across" is needed. Any understanding should be considered provisional, is always contested, and emerges from dialogue and debate between implicated actors [11].

Contextual and performative factors impact the benefits of transparency measures. The perceived trustworthiness of organizations deploying AI mediates the perception of transparency communications between controllers and subjects [48, 103]. Thus, we should work towards a *relational* understanding of transparency, focusing on what makes transparency communications meaningful and trustworthy in subjects' eyes [103].

While human-to-human explanations are socially situated, machine explanations are directed at its technical internals [93]. "Social transparency" of AI systems offers accounts describing *who* did *what* with the AI system, *when* and *why* they did what they did. Such explanations are more holistic and make human-AI assemblages more concrete [93].[6] Poursabzi-Sangdeh et al. [282] also argue against absolute measures of model interpretability and for a relational, contextual approach based on observation of actual behavior.

What can we conclude from the preceding? Full transparency is infeasible for epistemological, ethical, and practical reasons. Explanations should be designed for particular audiences and contexts. A contextual, relational, and sociotechnical approach is necessary for achieving meaningful transparency. The empirical account of the relationship between explanations, understanding, trust, and control is complicated and muddled. The least we can say is that transparency alone is insufficient for control.

Knowledge Gap and Research Question #1

We can now formulate the first knowledge gap. Public AI systems play a role in policy execution and affect citizens. In contrast to users and consumers, citizens are directly or indirectly impacted by such systems whether they choose to or not. Experts design transparency interfaces that provide explanations to support citizens' understanding, trust, and control. Therefore it is crucial to address how transparency mediates the relationship between experts and citizens. We can improve this relationship by understanding the different conceptions of transparency held by the major stakeholders involved.

6. Notice how these socially situated explanations resemble the justifications advocated for by Fine Licht and Fine Licht [106] and Henin and Le Métayer [146].

This leads us to RQ1: *What are the diverging conceptions of transparency between experts who design, develop, and govern public AI systems and citizens who use those same systems?*

1.2.2 Contestability

As seen in Section 1.2.1, if accountability is the goal, transparency alone is insufficient [156]. Improvements to the current accountability regime require, amongst other things, means for subjects to address and redress problems; and the availability of experts who can help challenge decisions [156].

An AI systems' legitimacy can be decomposed into three components: (1) input, (2) throughput, and (3) output. Threats to each type of legitimacy can be mitigated with institutional arrangements: legal structures, civic participation, and monitoring all play a role. Transparency mainly addresses throughput legitimacy. Civic participation in design and monitoring aids with input legitimacy. A *right to human intervention* aids with output legitimacy [138].

In cases where transparency is challenging to attain, for all the reasons already discussed, Walmsley [354] claims that *contestability* is a viable and acceptable alternative. There is no need to understand exactly how a decision was made to contest it. Contestability can take the form of civic participation built into the development phase, a human-in-the-loop during decision-making, or a feedback loop from decision subjects back to ongoing system development. Even if we cannot fully understand how a particular output was generated, we can at least challenge it [354].

Henin and Le Métayer [146] consider contestability as a way to require controllers to provide not just explanations (factual accounts) but also *justifications* (normative accounts) because contestations are arguments for a decision being not merely incorrect but also *undesirable* in some way. In this way, contestability contributes to accountability—the explanations and justifications are the accounts provided by controllers—and ultimately, legitimacy.

Conceptualizing Contestable AI

Research on contestable AI has been expanding, highlighting its significance in safeguarding against flawed and unjust automated decision-making by emphasizing human involvement and fostering adversarial discussions between decision subjects and system operators [9, 55, 146, 153, 298, 341].

Contestability can be viewed as humans questioning machine predictions, allowing human intervention to rectify potential machine errors [153, 350].

It can be described as a blend of human and machine decision-making, emphasizing its role in procedural justice and enhancing perceived fairness [222, 341, 366]. The practice of "contestability by design" stresses human intervention retrospectively and in the AI development processes [9]. Contestability transcends mere human intervention, demanding a dialectical interaction between decision subjects and human controllers [298]. A system's legitimacy is compromised without contestability, which demands *justifications* in addition to explanations [146]. Implementing contestability features in practice will require thoughtful consideration of needs, values, and context [223].

We conceptualize contestable AI as *systems that are open to human intervention throughout their lifecycle, emphasizing a dialogical relationship with decision subjects*.[7] Contestations can be leveraged for continuous system improvement (cf. Chapter 3). We also emphasize the relevance of participatory policy-making approaches and the need to monitor contestations for systemic flaws (cf. Chapter 4).

Knowledge Gap and Research Question #2

This brings us to the second knowledge gap. Work on contestability has, for the most part, focused on principles rather than system features and design and development practices. This focus on principles limits the ability of practitioners to take action [247]. We lack a complete and coherent description of the *actionable* sociotechnical system properties that make AI systems contestable.

From this we can derive RQ2: *What socio-technical features and practices contribute to AI system contestability?*

1.2.3 Public AI

Recall that I define public AI as the application of adaptive data analysis and processing to enhance, assist, or automate decision-making in the public sector [265, 326] (cf. Section 1.1.1).

Transparency is often considered a key to good governance, an idea encapsulated by Jeremy Bentham's claim that 'the more closely we are watched, the better we behave.' However, measures aimed at enhancing transparency in government often result in stricter and more centralized management of information rather than promoting openness in governance [159].

7. Here I write "we" because this conceptualization is the result of the collaborative work done by my co-authors and myself, reported on in Chapter 3.

Studying transparency in the context of public sector AI, Veale et al. [346] find a disconnect between research and practice. Research seeking to create impact should focus on in-context studies. Practice requires usable transparency tools for identifying risks and including domain knowledge in decisions. These tools should be aimed at both managers and frontline civil servants [346]—also called "street-level bureaucrats" [210, 211]. Inclusion of domain knowledge is of particular importance because of the need for the preservation of room for individual discretion in public sector human-AI decision-making [8, 34, 235, 276, 365].

Brown et al. [48] show how trust in public sector AI is mediated by subjects' trust in the organization deploying it. They propose several interventions to increase trust through transparency and communication, notably for our purposes; these include *the support of positive communicative relations between civil servants and subjects*.

Katell et al. [180] argue that if accountability is the goal, one should co-develop interventions with affected communities in context. This way, subjects are empowered, and interventions are more likely to address real needs. Empirical work shows that many helpful interventions turn out to be non-technical. A co-design approach ensures that problems are not framed a priori as solvable through more data—what Morozov calls "tech solutionism" [248]—but instead, the departure point is to ask if a particular system should be used at all [180]—sometimes referred to as a "politics of refusal" [367].

We need means to ensure the accountability and legitimacy of government use of AI that do not rely primarily on high degrees of transparency. Any work towards this should be conducted in context and with affected stakeholders. Understanding and supporting discretion is essential because it enables increased responsivity on the part of human-AI decision systems and makes the process of executing policy more practically feasible. Interventions that hold promise are relational, connecting developers with civil servants and affected subjects and improving dialogue between them. Affected communities should be brought into the development process of interventions.

Knowledge Gap and Research Question #3

This brings us to the third knowledge gap. Given the claim that contestability is a more effective alternative to transparency for holding public sector AI to account, it is necessary to understand its implementation's potential challenges. Several perspectives can be taken to explore these challenges. Following Veale

et al. [346], I choose to explore the issue from inside a public administration context in close collaboration with public servants.

This leads to RQ3: *What are the challenges facing the implementation of contestability measures in public AI?*

1.2.4 Agonistic Pluralism

The final subject to review is *agonistic pluralism* [249–255], the political philosophy that, to a large extent, animates the contestable AI field [66, 149]. I briefly touched on this topic in Section 1.1.4. Here, I will flesh out its conceptualization further and discuss how it has been invoked and applied in work on the design of AI.

Agonistic pluralism is a democratic model that prioritizes productive conflict over consensus, recognizing that a fully pluralistic society is unattainable but asserting that conflict is vital for maintaining diversity and preventing homogeneity. It supports open spaces for debate and challenges to power structures, emphasizing conflict as a fundamental aspect of society. This approach promotes ongoing debate over fixed values to encourage diversity and reveal power imbalances. It considers identities to be formed through political interactions. It aims to convert hostile relationships into a contest between legitimate political opponents, contrasting with models of democracy that focus on consensus and deliberation [75, 216, 304].

The concept of agonistic political design in AI systems is introduced to address and confront power relations [75]. It is argued that adversarial design techniques can democratize technology development by embracing agonistic principles [272]. Viewing AI through an agonistic lens reveals that these systems are always involved in contested spaces, with algorithmic decision-making representing a temporary balance of power [66]. Agonistic approaches to AI development allow society to make informed choices about adopting and integrating AI technologies. They enable individuals to contest or opt out of computational systems [148] and demand a more inclusive form of participation that respects the potential for conflict and power imbalances [289]. This perspective reframes AI not merely as a passive entity in politics but as an active political space, challenging binary views of AI as either wholly emancipatory or suppressive [89]. Finally, it suggests that AI safety should be pursued through "machine politics," promoting agonistic debate as not just a means to an end but as the primary objective itself [76].

Knowledge Gap and Research Question #4

We can now describe the fourth and final knowledge gap. Given that agonistic pluralism animates the contestable AI field, and given that design, on one level, deals with the framing and reframing of the concepts that shape how problems are understood and solutions are articulated, it is necessary to construct and communicate a clear description of the guiding concept that embodies agonistic AI design intended for the audience of public AI designers.

This leads us to RQ4: *What strategic guiding concept best complements contestable AI prescriptions of a more tactical nature?*

1.3 Approach

In this section, I describe the general approach taken to conduct the research reported in this thesis, which I frame as *constructive design research*. I discuss this research's methodological commitments, conception of knowledge, design methods, and methods of analysis.

1.3.1 Methodological Commitments

My epistemological and ontological commitments are contextualist [137, 147, 169, 225] and critical-realist [115, 132]. The consequences of these commitments for the approach taken are as follows.

A contextualist epistemology sits between positivism and constructionism. It has as its central metaphor humans acting in context. It does not assume a single reality and considers knowledge to emerge from context. Such knowledge reflects the researcher's position. Knowledge is localized, situated, and therefore always provisional. Despite this, contextualists are still interested in understanding truth, even though they hold that no single method can get to the truth. Knowledge will be true (valid) in certain contexts.

A critical-realist ontology sits between realism and relativism. It holds that there is a real and knowable world, but it sits 'behind' the subjective and socially located knowledge we can access. Knowledge is socially influenced and reflects a separate reality to which we have only partial access. For knowledge to make a difference, some authentic reality must exist. It is this external reality that provides the foundation for knowledge.

What this means for my approach is that data is always generated in context, and the knowledge generated through analysis is always presented and situated in potential application contexts, taking into account my positionality.[8]

1.3.2 Constructive Design Research

I frame the overall approach as constructive design research in the "field" and "showroom" modes [193, 194, 198]. Constructive design research is where the making of things—products, services, systems, spaces, media—occupies a central place and becomes the primary vehicle for knowledge generation.

The *field* mode follows design experiments as they move through society. It seeks to understand how humans make sense of things in context. This sense-making is usually done through creating prototypes and evaluating them in the field with humans. Crucially, the aim is usually not, as in design practice, to evaluate the prototype itself. Instead, the prototype is an instrument for generating data about the phenomenon of interest. Such field experiments can also include group interactions and respondents' participation in creative acts.

In the *showroom* mode, design is performed as an act of creative and artistic experimentation and societal critique. It focuses on uncovering, debating, and reinterpreting matters of concern. Artifacts in this mode frequently project contemporary developments in science and technology and project them into the future to explore alternative pathways and elicit potential future social consequences. To engender a critical attitude in audiences, design researchers in this mode often use a tactic of estrangement or defamiliarization. Researchers are typically concerned with matters of form over matters of use, the stories their designs tell, and the stories they provoke in audiences.

This thesis' approach sits somewhere in between these two modes. Where the focus lies between these two modes varies from study to study. We take designs into the field and are primarily interested in questions related to how humans make sense of and act in the world.[9] However, we also bring critique and speculation into the creation of artifacts and desire for them to stand on their own and speak for themselves as objects of design.

8. The exception to this is formed by Chapter 3, which reports on a systematic literature review. Here the aim is to develop practice-oriented design theory that can aid subsequent in-context studies.
9. In this case, the "we" refers to my co-authors, collaborators, and myself. Team composition varied from study to study. Full credit is included in the individual chapters.

1.3.3 Knowledge

I seek to contribute to design research and practice by creating new *generative intermediate-level design knowledge* [161, 214]. *Generative* means knowledge offers the seed for a design solution with particular qualities without fully prescribing its shape. *Intermediate-level* means knowledge occupies a continuum between specific instances of designed artifacts and generalized theoretical knowledge.

When I engage in design in the interest of knowledge development, instead of improving a use situation, I take inspiration from *concept-driven interaction design* [321]. This approach seeks to "manifest theoretical concepts in concrete designs" and complement user-centered design approaches in HCI [321]. I have taken a concept-driven approach mainly in the Contestable Camera Cars concept video (Chapter 4) but also, to some extent, with the Transparent Charging Station prototype (Chapter 2).

In thinking about how designers frame and solve problems, I also rely on Schön [302]'s notion of *generative metaphor*. These are guiding concepts that influence perception and understanding of the world. Generative metaphor functions by transferring perspectives between domains. The resulting perception affects decisions and actions. We make use of generative metaphor, in particular in Chapter 5 to construct from theory a guiding concept for contestable AI design, as well as to analyze concept designs generated by workshop participants (cf. Analysis, below).[10]

1.3.4 Design Methods

The studies reported on in this thesis make use of several construction techniques: prototyping [129, 208, 317, 318, 357], speculative design [16, 33, 110, 119], and information design [337, 358].

Prototypes in the context of design research have been defined as "things we make to find out things" [317]. In design practice, prototypes are generally intended to demonstrate and validate ideas. When used in design research, they function as instruments, in the way already mentioned when discussing the field mode of constructive design research. We used prototyping when co-creating the Transparent Charging Station (Chapter 2). A non-functional prototype of a transparency interface that could be added to existing public charge points

10. The "we" here is deliberate because my co-authors and I did the construction and analysis in collaboration. Full credit is provided in the respective chapter.

was evaluated with EV drivers in the wild to understand their conceptions of AI transparency better.

When we make use of forms of storytelling to convey design ideas, my collaborators and I take inspiration from *speculative design*, design fiction, and other forms of design futuring [16, 33, 110, 119]. The logic of this practice has already been covered when discussing the showroom mode of constructive design research. Speculative design is typically framed as a design practice that asks questions rather than solves problems. Because of this, the audience's thoughtful engagement is considered its own form of success [119], what in the words of Haraway [141] is described as "staying with the trouble." Others have rejected this dichotomy and instead claim speculative designs are distinguished by so-called "parafunctionality," a type of design where function is used to encourage reflection on how products condition behavior [227–229]. The Contestable Camera Cars concept video (Chapter 4) was created in this fashion and fits best with the notion of para-functionality. Near-future depictions of contestable camera cars were used to encourage critical reflections on the viability of those ideas in participants, civil servants who work with AI.

To communicate design knowledge to practitioners, my collaborators and I use forms of visual communication, particularly a type of *information design* that has been described as "visual explanations" by Tufte [337]. Visual explanations are "*pictures of verbs*, the representation of mechanism and motion, of process and dynamics, of causes and effects, of explanation and narrative" [337]. Such infographics are suitable for depicting systems-oriented knowledge and are especially beneficial for practitioners who often rely on visual aids [358]. We used this technique to create the Contestability Loops for Public AI infographic (Chapter 5). We created a large, illustrated diagram that conveys a part of the provisional design framework so that practitioners can use it to guide early-stage concept design work on public AI systems. This infographic was evaluated together with professional designers in a series of workshops.

Finally, all artifacts employed in the studies reported here were constructed in collaboration with professional designers. These collaborations allowed us to achieve a high degree of creative design excellence (or "polish") without sacrificing scientific quality. The back-and-forth these collaborations engendered between concerns of design practice and design research further contributed to artifacts with a high degree of coherence between form, function, and theory. Those dialogues also contributed to further sharpening study aims, approaches, and analysis. The chapters clearly credit collaborators and provide accounts of their contributions to the artifacts in question.

1.3.5 Analysis

Throughout these studies, for data analysis, the chief method used was *reflexive thematic analysis* [39–41, 43–45], a qualitative approach to data analysis with its origins in qualitative psychology. It is compatible with my methodological commitments, accessible, with a high degree of flexibility, allowing easy adaptation to a design research context. Reflexive thematic analysis accommodates a range of questions, data generation methods, and sample sizes. Furthermore, its results tend to be accessible to a lay audience, which fits well with a participatory, action-oriented approach.

As with other forms of thematic analysis, the method relies on procedures of coding qualitative data—usually text, such as interview and focus group transcripts, but other data types are possible. These codes are developed into themes through deductive and inductive forms of analysis. Coding can be done for manifest meanings—explicit statements in the data—or latent ones, meanings that underly the data.

Good reflexive thematic analysis considers the researcher as the primary "tool" and embraces subjectivity; aims for analysis that is strong (deep) rather than objective or accurate; uses collaborative coding not to achieve reliability but to enhance reflexivity; considers themes to be constructed *from* codes rather than to be *applied* to data; considers themes to represent coherent ideas rather than summaries around a topic; sees themes not as "emerging" or "discovered" but as actively constructed by researchers through deep systematic engagement with data; is underpinned by theoretical assumptions, which must be acknowledged; insists that researchers own their perspective; and conceptualizes analysis as an art, not a science—a creative practice within a framework of rigor [40].

Most of the analysis took a bottom-up approach typical of conventional reflexive thematic analysis. The themes were constructed from the data in an open fashion, with little theoretical guidance up-front (this applies to Chapters 2, 3 and 4). The final study (Chapter 5) took inspiration from critical realist approaches to reflexive thematic analysis, which allow for the development of themes using apriori theoretical frameworks and can yield causal explanations [118, 359].

Two more elements shaped the analysis, particularly in the final study, reported on in Chapter 5: the aforementioned generative metaphor [302], and *annotated portfolios* [35, 120, 214].

Schön's generative metaphor was used to construct a guiding concept for design and as a theoretical lens to trace concept designs via their shared 'mechanisms' to several metaphors latent in the design space. There is precedent in

HCI design research related to AI for using generative metaphor as a constructive and analytic tool. For example, Dove and Fayard [84] uses the technology as a monster metaphor to frame and reframe how designers relate to ML as a design material; Murray-Rust et al. [259] catalog existing metaphors in AI discourse that they deem inadequate and propose a range of alternatives; Benjamin et al. [26] describes a metaphor-driven research-through-design project that uses the metaphor of entoptic phenomena to conceptualize the mediation of reality by prompt-driven AI image generation; and finally, Nicenboim et al. [261] use the metaphor of AI as home-grown organisms to explore productive misunderstandings of smart home speakers.

The analytic means by which those mechanisms from concept designs were constructed was, in turn, inspired by annotated portfolios [35, 120, 214]. The concept of annotated portfolios involves choosing a set of designs, displaying them in a suitable format, and supplementing the design displays with short written explanations. Bowers [35] and Gaver and Bowers [120] describe this approach as a way to communicate design research, which is familiar to both designers and artists [214]. In this adapted approach, participants' sketches of concept designs are supplemented by verbal descriptions, and the family resemblances typical of annotated portfolios were the main focus of coding and theme development.

1.4 Chapter Overview

This is a paper-based PhD thesis. Each chapter of the thesis contains either a peer-reviewed journal article or a peer-reviewed conference paper. The venues these are published in cover interdisciplinary work on the relationship between AI and society, the philosophy of AI, human-computer interaction, and design.

In Chapter 2, my co-authors and I investigate the diverging conceptualizations of AI transparency by experts and citizens in the context of smart electric vehicle charging.[11] We show that absent means of control, people find transparency measures irrelevant and burdensome. These findings illustrate the need for contestability.

In Chapter 3, we conduct a systematic literature review of contestable AI and construct a provisional design framework from its findings. This framework describes five system features, their relationship to the major human-AI system

11. My co-authors are credited in the respective chapter. In the remainder of this section, I use "we" to refer to my co-authors and myself.

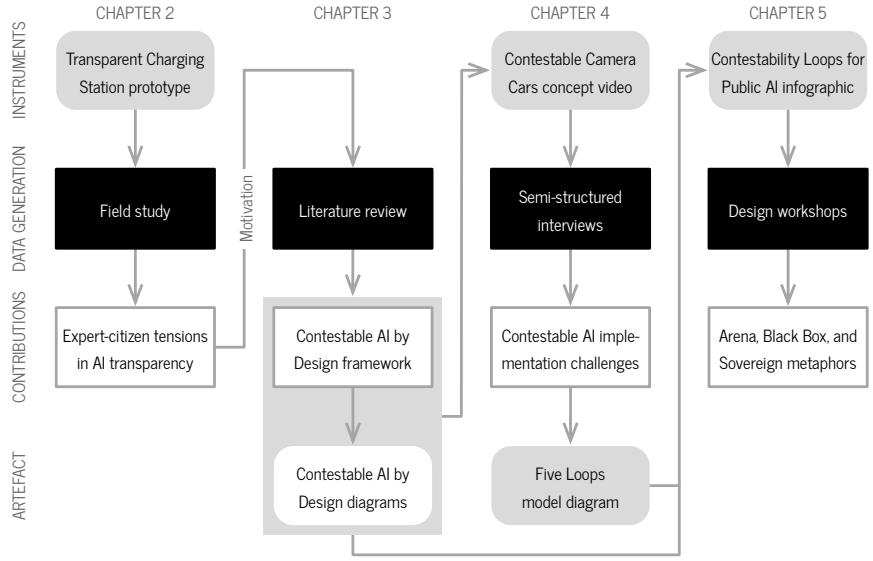


Figure 1.4
Diagram showing how the chapters, design artifacts, data generation methods, and contributions relate. Chapter 2 describes the creation of the Transparent Charging Station prototype and analyzes the data generated with its help in a field study. This study motivated the remainder of the PhD project. Chapter 3 describes the creation of the Contestable AI by Design framework employing a systematic literature review and the accompanying diagrams. Chapter 4 describes the creation of the Contestable Camera Cars concept video, which applies the framework to the case of camera cars. It analyzes data generated with its help in semi-structured interviews. This analysis furthermore yields the Five Loops model. Chapter 5 reports on constructing the Agonistic Arena metaphor and creating the Contestability Loops for Public AI infographic, which is also informed by the framework and the Five Loops model. It reports on data generated with the help of the infographic in concept design workshops paired with focus group discussions.

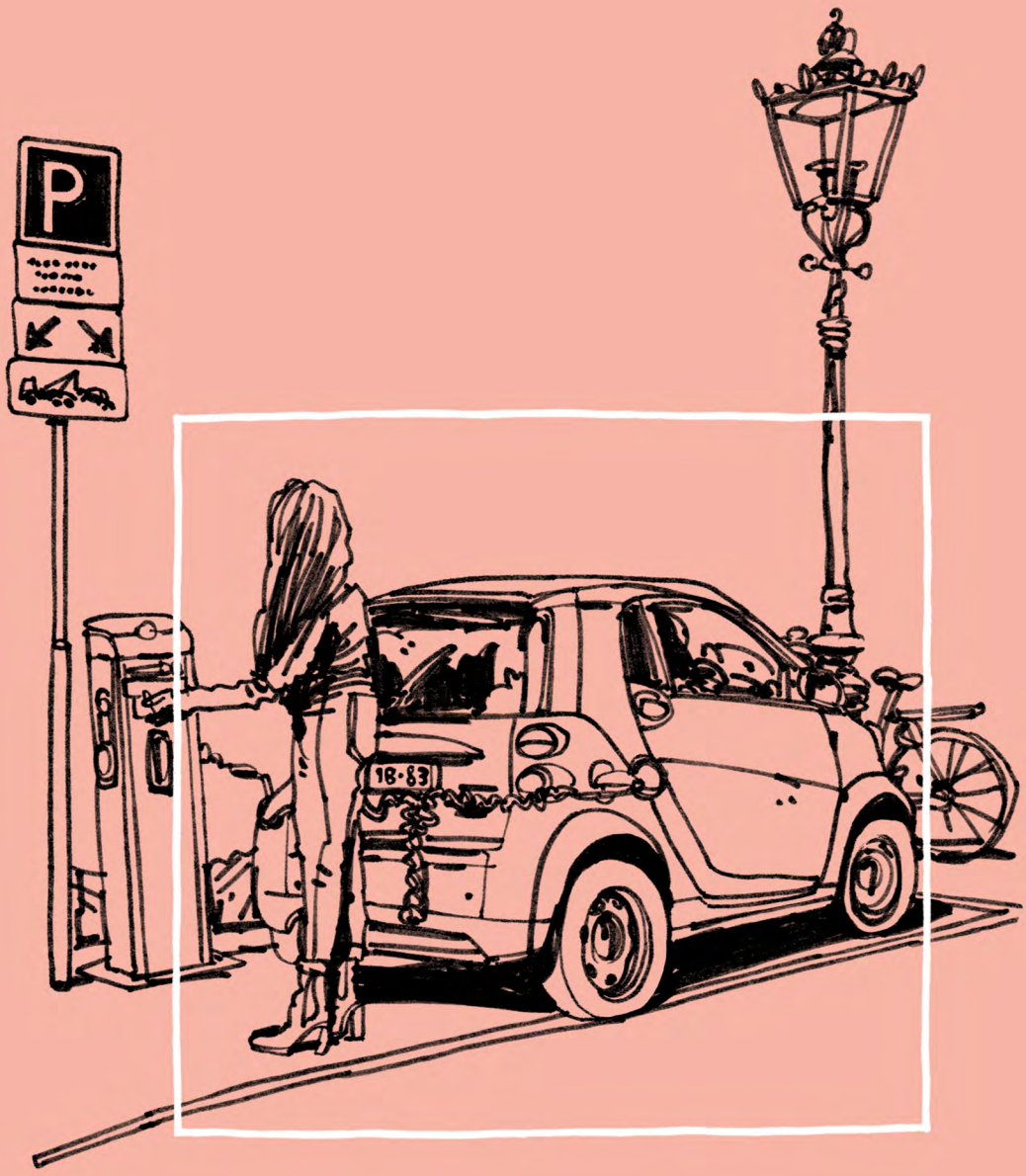
actors, and six practices and their relationship to the stages of a typical AI system development lifecycle.

In Chapter 4, we apply the design framework in the context of camera cars. We create a speculative concept video of a near-future contestable camera car. We use the video to conduct semi-structured interviews with public servants employed by the city of Amsterdam who work with AI. The findings explore various challenges facing the implementation of contestability in practice.

Finally, in Chapter 5, we use theory to construct the generative metaphor of the Agonistic Arena. We create an infographic that communicates this metaphor

and parts of the framework to a professional design audience. We conduct workshop focus groups at design agencies in which the infographic is applied to the case of a vacation rental fraud risk model. The findings report on the mechanisms shared across participants' concept designs and three competing metaphors to which these can be traced.

In the concluding discussion, I answer the research questions, reflect on the implications of the findings, and consider limitations and future work. Figure 1.4 presents an overview of the relationships between chapters, cases, design artifacts, data generation methods, and contributions.


P
18·83

Chapter 2

Tensions in Transparent Urban AI: Designing a Smart Electric Vehicle Charge Point

Citation:
Alfrink, K., Keller, I., Doorn, N., and Kortuem, G. "Tensions in Transparent Urban AI: Designing a Smart Electric Vehicle Charge Point." In: *AI & Society* 38.3 (Mar. 2022), pp. 1049–1065. DOI: 10/gpszwh

Abstract:
The increasing use of artificial intelligence (AI) by public actors has led to a push for more transparency. Previous research has conceptualized AI transparency as knowledge that empowers citizens and experts to make informed choices about the use and governance of AI. Conversely, in this paper, we critically examine if transparency-as-knowledge is an appropriate concept for a public realm where private interests intersect with democratic concerns. We conduct a practice-based design research study in which we prototype and evaluate a transparent smart electric vehicle charge point and investigate experts' and citizens' understanding of AI transparency. We find that citizens experience transparency as burdensome; experts hope transparency ensures acceptance, while citizens are mostly indifferent to AI, and with absent means of control, citizens question transparency's relevance. The tensions we identify suggest transparency cannot be reduced to a product feature but should be seen as a mediator of debate between experts and citizens.

2.1 Introduction

Digital technologies such as big data, sensor networks, and artificial intelligence (AI) are becoming increasingly important in the control of urban infrastructure and public administration more broadly [61, 67]. However, it is now widely recognized such AI systems may lead to unfair outcomes, even if they have been

designed with the best intentions [98, 286]. These concerns have prompted researchers, governments, and civil society groups to formulate ethical principles for the deployment and use of AI, emphasizing values such as transparency, fairness, and accountability [174, 242, 335]. Likewise, some cities have started to embrace a digital rights agenda and are formulating principles and policies to govern public AI systems (e.g., [331]).

Many ethical and policy frameworks see *transparency* as an important prerequisite for ensuring fairness and public acceptance [46, 322]. Empirical research in human-computer interaction (HCI) has focused on identifying which forms of user interface-level transparency are most effective for increasing user understanding and trust [1]. In this HCI research, transparency is typically framed as a form of objective knowledge that empowers people to make informed choices about how best to use and govern AI systems. However, researchers have started to point out theoretical and practical limitations of the transparency ideal [11], and the importance of considering the human experience of AI transparency [10, 344]. What is more, in the case of *public* AI systems, such as those controlling urban infrastructure, i.e., "urban AI," the relationship between users and those who design, develop, and govern systems is different from commercial settings: These systems effectively *enact policy* [187, 307], and users are not simply consumers, but also *citizens* who are entitled to democratic control over policy, AI-enacted or otherwise.

Therefore, our aim is to examine the degree to which transparency-as-knowledge is a suitable concept for urban AI systems in both an empirical and critical way. We contribute to the ongoing discussion of transparent AI by investigating diverging conceptions of transparency between those who design, develop, and govern urban AI systems (hereafter "experts") and users of those same systems ("citizens").

We focus on smart electric vehicle (EV) charging as an empirical ground for studying transparency in urban AI systems. Smart EV charging serves as a useful example of how urban AI shapes the lived experience of cities and of city-making itself. Smart EV charging facilities augment and mediate both public spaces and travel spaces. In this context, as well many stakeholders consider transparency an essential ingredient for ensuring public acceptance [77, 99, 224]. Using a practice-based design research approach [194], we collaborated with commercial companies and the municipality of Amsterdam to prototype and evaluate a *transparent* smart EV charge point which provides EV drivers with explanations of smart charging decisions.

The significance of our findings lies in shedding light on several tensions between motivations experts have for providing transparency, such as social

acceptance, and attitudes and expectations citizens have towards urban AI systems, such as indifference or a desire for control.

In what follows, we first briefly provide context on smart EV charging and the Amsterdam design project that formed the basis for our empirical work. We then summarise work on transparency in HCI design research, philosophy of technology, and the social sciences of big data and AI. Subsequently, we describe the field study we undertook with the design project prototype. Following this, we offer six narrative themes to capture our findings with regard to expert understanding and citizen experience of transparency. In the concluding discussion, we contextualize these findings in light of the literature and examine the main points of tension between expert understanding and citizen experience.

2.2 Background and Motivation

2.2.1 Smart Electric Vehicle Charging

Many cities see electric mobility as a key way to improve efficiency and equity of the flow of goods and people and to reduce negative externalities, including air pollution and climate change [122]. However, in OECD countries, there are indicators that electric grid capacity is not sufficient to support the growing number of EVs [257]. In general, this is not an issue of overall energy availability but of limited grid capacity [164]. This concern is especially relevant for local distribution grids in cities and neighborhoods where EVs are particularly prevalent and where charging sessions are clustered around peak times. If demand for charging exceeds supply, not every vehicle can be charged, and choices need to be made: who will be charged first, and who will last? For this reason, energy network providers have started to deploy "smart charging" solutions, which make the timing and capacity of EVs dependent on factors such as grid capacity, electricity demand, and availability of renewable energy. Smart charging allows for dynamic management of demand by curtailing the rate and amount of electricity EVs can charge when connected to a charge point [117, 234, 243, 355]. The use of AI in governing the grid and charge points makes it possible to increase the number of EVs by more than 60 percent without having to upgrade physical grid infrastructure [269].

Of course, EVs are not an unambiguously positive development, nor should the transition to EVs be considered inevitable. In fact, EVs are a contested subject involving many social, political, and ethical debates. To name but a few concerns: EVs may perpetuate existing car culture, increased electricity needs may not be met by renewable sources, battery production depends on the exploitation

of limited mineral resources with adverse social and ecological consequences, and EV battery recycling itself can cause pollution [270].

Smart grid solutions may reinforce and accelerate practices producing energy demand peaks rather than contributing to more sustainable ways of living [323]. A focus on solving the problem of demand also distracts from rethinking everyday practices requiring energy in the first place [324, 325]. In other words, smart EV charging can be seen as a form of "technological solutionism" [248], where social ills are framed as problems to be fixed by means of technology while avoiding structural change [113].

In any case, smart charging solutions significantly alter the EV charging experience: EVs may charge slower than expected; drivers may be disadvantaged by receiving less electricity or slower charging rates than other drivers, even if both cars are plugged in at the same time and charge point. It may also have unexpected side effects, such as some neighborhoods receiving less electricity than others. In short, the use of AI makes EV charging less predictable. From the perspective of experts, this threatens social acceptance. Transparency promises to contribute to people's understanding of and trust in smart EV charging systems.

2.2.2 'The Transparent Charging Station'

As of July 2021, the city of Amsterdam operates 2503 charging stations, or 4974 charging points [124]. Of these, at the time of our study, 452 stations (904 points) were part of a smart charging system called Flexpower [125]. This system increases charge speed when solar energy is available and decreases speed around peak times when the grid is used more intensively.

Prompted by rising public concern about the risks of the Internet of Things and AI, in 2016, electric grid operator Alliander[1] and EV charging knowledge institute ElaadNL,[2] commissioned a design study from design agency The Incredible Machine[3] to develop ways of making smart charging transparent for EV drivers. The outcome was the *Transparent Charging Station*, a speculative design prototype of a smart charge point using a video game metaphor for visualizing automated charging decisions [338]. A key aspect of the *Transparent Charging Station* is the use of priority schemes: for example, shared EVs would get priority access to charge faster, sooner, and more than non-shared private vehicles. The design study received significant public interest but also raised questions about

1. https://www.alliander.com
2. https://www.elaad.nl
3. https://www.the-incredible-machine.com

the meaning, viability, and utility of transparency in the context of a street-level public service.

A follow-up project, *UI for Smart EV Charging*, was initiated in 2019 by the same knowledge institute and design agency, which were joined by the municipality of Amsterdam and the Amsterdam Institute for Advanced Metropolitan Solutions.[4] The aim was to develop a prototype transparency interface inspired by but distinct from the *Transparent Charging Station* speculative prototype. The project built on the newly formulated digital agenda of the city of Amsterdam entitled *A Digital City for and by Everyone*, which lays out values and ambitions for a "free and inclusive digital city" in which the digital rights of all residents are protected [123]. This design was aimed at a solution compatible with existing Flexpower charge points in an effort to further study the technical feasibility, usefulness, usability, and desirability of transparency provided through a screen-based user interface. The first and last authors agreed to become part of this project group to consult during the design phase and lead the evaluation of the design solution. Simultaneously, we pursued our independent research agenda into the varying conceptions of transparency by major direct stakeholders involved in urban AI projects.

2.3 Related Work

Transparency is a widely held and discussed moral and political value, especially in settings where informed consent, accountability, and deliberation are emphasized. In the context of AI, in particular, when developed using machine learning (ML), transparency commonly refers to the visibility and accessibility of information related to a system's functioning. The opacity of AI systems can stem from a variety of sources: deliberate secrecy by system developers and operators, lack of technical literacy of the observer, or technical properties of systems themselves [51]. In particular, transparent AI aims to provide *explanations* of model behavior. Such explanations can be arrived at by developing models that are *interpretable* by humans, for example, because they are rule-based. When models are developed with techniques producing opaque or "black box" models resisting human interpretation, explanations can still be produced in a post hoc fashion by means of a supplemental explanation model [182, 371].

In debates around the social and ethical ramifications of AI, transparency has quickly become a central if contested notion. Many view transparency as a

4. https://www.ams-institute.org

desirable value, either for moral reasons or because it aids understanding and increases trust. Others point out AI systems resist straightforward explanations due to their sociotechnical nature. They warn against how transparency shifts responsibility from system developers to users.

Surveying the literature on big data and AI in HCI design research, philosophy of technology, and interdisciplinary work, we can identify this same emphasis on the relationship between transparency, understanding, and trust. There is also a growing body of critical work exploring transparency's limits.

2.3.1 Transparency, Understanding, and Trust

The main vehicle for creating transparency of AI systems on the level of user interfaces is through so-called "explanations," informational and/or interactive elements communicating some aspect of an AI system's workings. Various kinds of explanations can contribute to people's understanding of an AI system [284].

Explanation completeness and soundness impact the fidelity of end users' mental models. Explanations with a high level of completeness have the lowest perceived cost and highest benefit. However, this favorable cost-benefit perception does hinge on users being able to adjust system behavior [200]. Furthermore, when users feel they are able to form an adequate mental model from simply interacting with systems, explanations are less likely to be considered beneficial because they take attention away from primary tasks [50].

Increasing transparency by providing explanations can improve people's trust in AI systems [30, 97, 188]. User literacy of algorithmic systems may mediate the degree to which explanations increase trust [309]. There does not appear to be a single best way of explaining a system to increase trust [30]. In some cases, trust only increases as a result of explanations when user expectations have been violated by system behavior [188]. There may also be a bell-curved relationship between information amount and user trust. Providing too much information can actually *erode* trust [188]. Trust does not appear to be impacted by people's objective understanding of systems nor by the form of explanation used [60]. There is some evidence explanations need not even be *truthful* to increase user trust [94].

User trust may also be impacted significantly by their attitudes to the larger systems that form the context of automated decision-making. For example, looking at the application of AI in child welfare services, Brown et al. [48] find people's distrust of non-automated systems increases their discomfort with AI.

A tension exists between making people aware of AI's functioning and preventing them from developing behaviors at odds with system developer goals. A

level of obfuscation is necessary to prevent bad actors from gaming the system, whereas a lack of transparency reduces people's sense of control and makes them unsure about how their behavior might impact outcomes [10, 96, 173].

2.3.2 Critiques of Transparency

In the philosophy of technology and interdisciplinary work on the social implications of big data and AI, critical efforts have explored the limitations of the transparency ideal.

The language of transparency suggests we are removing things obscuring our view, while in fact, transparency requires active production of information [237]. However, more data does not necessarily lead to better understanding. In our current age, it is not a lack of information but a sheer abundance of data obscuring our understanding [53]. Furthermore, when we strive to make automated decisions explainable, we should be wary of the distinction between appearing transparent and actually being transparent. The latter requires *actionable* information, that is to say, information humans can use as a resource for their own decision-making [344]. Publishing (non-actionable) information in an effort to merely appear transparent can be a form of "tokenism" or "engagement theatre" in that it does not actually increase democratic control over urban AI systems [178, 246]. Indeed, large tech companies use transparency initiatives at least in part to stave off government regulation [133]. Another risk of focusing on transparency is that it makes us less likely to consider if we want an AI to determine a particular aspect of our lives at all [71].

Transparent AI is often treated as an issue best dealt with behind closed doors by experts. AI presents challenges for traditional HCI design in general [158] and participatory design approaches in particular [37]. However, a small but growing body of work seeks to bridge the gap between advocating for abstract principles and supporting design choices situated in context [3] and opening up AI development processes to a broader range of stakeholders [196]. In this way, users and citizens gain control over ways in which transparency is implemented in particular AI systems so they support their needs.

The transparency ideal can reinforce a neoliberal model of human agency, in which perfectly informed and fully consenting individuals make rational decisions that, in the aggregate, produce improved social outcomes [11]. This model reduces citizenship to one of consumer choice. If, on the basis of the information provided, a person disagrees with a system's functioning, they are expected to defect to a competing but sufficiently equivalent service. In the language of Hirschman [154], this is the "exit" option. The alternative is

"voice": expressing disagreement in order to effect change. Hirschman suggests we focus on the latter because making space for and responding to feedback increases "loyalty." Since we are dealing with public AI systems, relying on exit alone is problematic because citizens should have a say in the operation of these systems, and limiting participation to "voting with your feet" infringes on people's right to the city [56, 112, 295, 307].

Making AI system data and models visible is not the same as holding whole sociotechnical assemblages accountable. For this, it is necessary to see who has the power to change systems and to be able to experiment with changes ourselves [11, 81, 157]. The sociotechnical complexity of urban AI systems may exceed individual capacity, in which case it can only be understood collectively [166]. Transparency can also not account for cases in which a system's behavior deviates from design intent due to adversarial attacks [73]. There are also ways of increasing accountability that do not depend on transparency at all. One example is to introduce ways for AI systems to exercise "discretion," to diverge from the baked-in policy in cases where user dissatisfaction with system behavior is detected [8]. Another is to include means for decision subjects to *contest* AI system decisions, to make them responsive to requests for human intervention [9, 153, 298, 341, 354].

Empirical work in HCI indicates transparency through user interface-level explanations can contribute to understanding and trust. However, both understanding and trust achieved in this way are highly contingent and may not even be justified in the objective sense. At the same time, critical work points to the limitations of the ideal of transparency, often questioning the motivations of system designers and developers, pointing out how their understanding and valuing of transparency may not be the same as that of users. When conflicting mental models and values are glossed over, design processes and outcomes are likely to suffer. Because public AI systems enact policy and in such settings, users are also citizens, the ways in which transparency interfaces mediate the relationship between experts and citizens should be considered together. We, therefore, argue it is necessary to improve our understanding of the varying conceptions of transparency by the major direct stakeholders involved.

2.4 Method

Our overall research approach is qualitative-interpretive. In order to investigate how experts understand transparency and how citizens experience a transparent AI system, we conducted what Koskinen et al. [194] describes as

a practice-based design research study in the "field" mode. In this approach, design methods such as interventions with prototypes in real-world settings are used to generate research data. In addition, we draw on participatory action research [181] for our active involvement in the industry project that produced the prototype. Our research consists of two main activities: (1) participation as design experts in an industry project to observe how experts conceptualize and implement transparent urban AI, and (2) evaluation in the field of the resulting design to understand how citizens experience urban AI transparency, as implemented in a transparent smart EV charge point. These activities we undertook as part of the *UI for Smart EV Charging* project (Section 2.2.2). Figure 2.1 provides an overview of the project structure.

Data collected consisted of project documents, field observations, and interviews. The first author was present at all meetings of the design team to observe and participate in the discussions. A reflexive field journal was kept, and documents produced during this phase, such as the design agency's project proposal and slide decks used during presentations (D1 through D10), were stored for future analysis. Analysis was performed using reflexive thematic analysis [45].

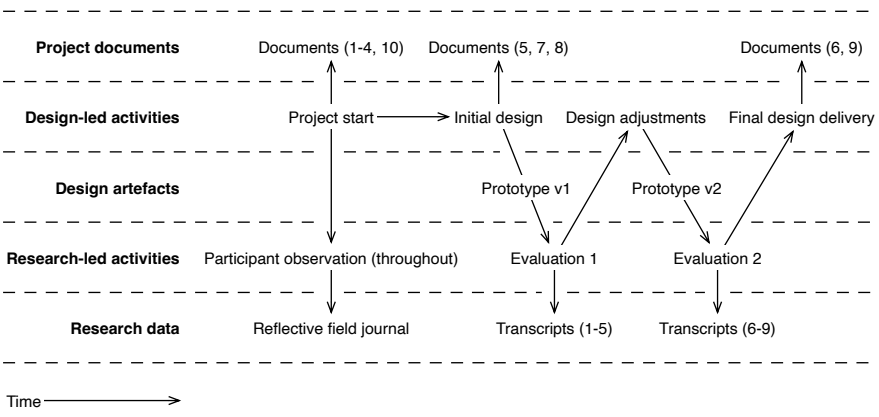


Figure 2.1
Overview of UI for Smart EV Charging project structure.

2.4.1 Design Process and Evaluation

The design and evaluation were done in sequence. The design phase was led by the design agency. Their starting points were their previous experience on the preceding speculative project, the project proposal drafted by the project partners to acquire funding (D10), and a requirements document developed by the project partners (D4). In the initial design phase, five EV drivers were interviewed to acquire insight into user needs (D7). A typical user journey for EV charging was mapped (D5). The first author spent a day at the design agency to ideate various approaches for the design. Following an exploration of various design options and feedback from the consortium, a final design was chosen. This design was developed into a high-fidelity, non-functional prototype.

For evaluation, a fast-charging facility centrally located in the Netherlands was selected as the field site. We set up the prototype next to fast charge points (Figure 2.2) and invited people who came to charge their cars to participate in the study. If they agreed, we went through an information sheet and consent form. In order to improve ecological validity, we did not provide participants with details on how the system operates beyond telling them we were testing the design of a "smart" EV charge point that adjusts speed based on a number of (unspecified) factors. Subsequently, we asked them to perform the task of charging their car using our prototype. While they did so, we invited them to think out loud and occasionally prompted them with open-ended questions. After completing the task, we followed up with a semi-structured interview to dig deeper into their experience with the prototype. All sessions were recorded using video and audio. Photographs were also taken. Furthermore, researchers took hand-written notes while observing. Overall, we conducted two rounds of one-day-long design evaluations: round one included five participants (P1 through P5; one female, four male), whereas round two included four participants (P6 through P9; one female, three male). Audio recordings of the evaluations were subsequently transcribed for further analysis. All quotes from participants and some from documents in this paper's results section were translated from Dutch by the first author.

2.4.2 Prototype

The prototype consists of a 1:1 scale cardboard replica of the charge points in use in Amsterdam. The signage on the stations is reproduced, and ports have been added for actual charge connectors to fit into. A 12.9-inch tablet is attached to the top of the charge point for the transparency interface to run on.

Figure 2.2
The design prototype was evaluated with EV drivers recruited on the spot at a fast-charging facility.

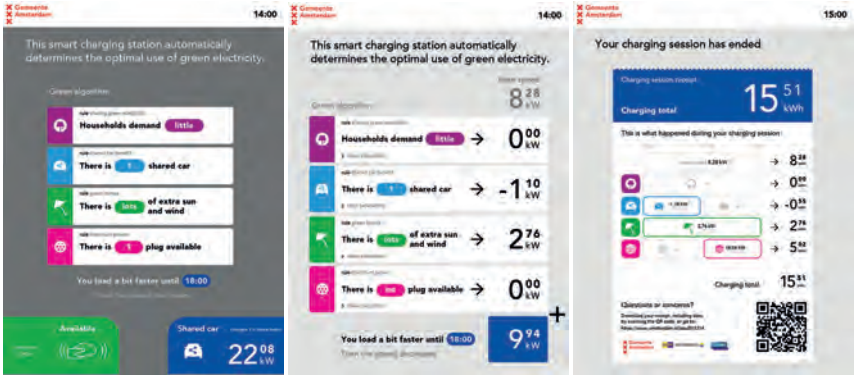
Figure 2.3 shows a selection of screens from the prototype (translated from Dutch by the first author). The basic structure consists of (1) an idle screen, (2) a screen shown once charging has started, and (3) a screen shown after charging has concluded. We distinguish two types of screen elements: those supporting the task of charging (e.g., a prompt to swipe a card to begin) and elements aiming to make the smart charging system *transparent*, i.e., explanations.

The prototype screens were created in a graphics package, and data reflecting the imagined scenario of use was added. The screens were collected in a presentation software file so it was possible to advance them using a concealed wireless remote control in response to participant actions. We created two versions of the user interface design, v1 and v2. V2 addressed some basic usability issues detected during the initial round of evaluations. These usability fixes aside, both versions of the prototype are identical.

‘Rules’ as Explanations

The main means of providing transparency is a set of elements that together list the “rules” governing system behavior. For each rule, its currently active state is displayed along with a short descriptive name. In v1, the other possible states are also immediately shown, and a few lines of additional explanatory text are included. In v2, each rule can be tapped to reveal a modal box that includes the additional text, the other possible states, and a graph or diagram offering a visual explanation. Once charging starts, each rule also includes an indication of how it impacts the charge speed. V1 uses amperes (A) as the indicator of charge speed (actually the unit of current). V2 instead uses kilowatt (kW) (the unit of power).

The screen displayed when a charging session is finished uses the conceit of a cash register receipt to show how much the user had charged in total, expressed in kWh. The receipt also shows any changes to each rule that may have occurred during charging while the user was away. A QR code and a unique URL for the charge session are also displayed, and some text next to it explains the code can be scanned or the URL accessed to receive a digital copy of this receipt.



(a) Idle (b) Session started (c) Session completed

Figure 2.3
Key screens of prototype v2.

2.4.3 Analysis

Two datasets were analyzed: dataset 1 compiles documents produced during the design project, and dataset 2 compiles data from prototype evaluation sessions

with EV drivers. Analysis was done using the qualitative data analysis software *Atlas.ti*. The data was first coded inductively by the first author. Codes were repeatedly refined and grouped into an initial set of themes. The second author independently coded a subset of the data, and a refined set of themes was jointly developed. The first author also checked the codes and themes with the commercial collaborators. Finally, the themes were once more condensed into the final, smaller, richer, and more narrative set presented in this paper.

2.5 Results

We generated six themes related to accounts of transparency in the data. There are two themes for experts derived from design project documents: (X1) truthful information produces transparency, and (X2) transparency enables fairness assessment. Four themes for citizens, derived from prototype evaluation session transcripts: (C1) transparency mediates concern; (C2) transparency is burdensome; (C3) transparency invites strategic behavior; and (C4) transparency evokes the desire for control. Almost all of the data was included in the themes.

2.5.1 Expert Understanding of Transparency

Theme X1: Truthful Information Produces Transparency

Experts talk about transparency as something created by providing *truthful* information about "automated decisions" (D4). The issue with these decisions is that they are *opaque*, hidden inside "black boxes" (D10).

> However, algorithms that currently control smart city objects are "black boxes": the public is affected by their decisions but does not know what factors are taken into consideration and how they are weighed against each other to reach a decision. (D10)

Here, we get a glimpse of what decisions an AI system makes: it weighs various factors against each other. However, throughout the documents, *decision-making* and *prioritization* are used interchangeably. We can also see that not only decisions but motivations for them must be made transparent.

Prioritization appears to produce dilemmas. Some people will lose, and others will win out in resource distribution.

> When a city service is scarce, prioritization is required. By using smart applications, cities need to prioritize beforehand and program them explicitly. It's possible to prioritize on the basis of target groups, like citizens, disabled, professionals, etc., shared vehicles, price, and time slots. (D10)

This notion of dilemmas connects to one of the driving motivations for pursuing transparent smart charging. It is not so much a moral imperative but a pragmatic one. The concern is that opacity threatens *acceptance* of EV driving and charging by citizens.

> Visibility [of] the automatic decision-making in the smart charging process can help the adoption of this new technology. (D10)

The conceptual metaphor used by stakeholders to describe *how* transparency is achieved is sometimes explicitly vision-based. Apparently, automated choices can be made *visible*.

> The Transparent Charging Station will provide insight into this by making the underlying choices of the algorithm visible on the display. (D1)

Theme X2: Transparency Enables Fairness Assessment

Transparency should enable users to determine if they have been fairly treated. Fairness assessment is impacted by design choices. For example, at one point during the design process, the design agency emphasized they had moved away from determining fairness by comparison.

> It is not about understanding fairness by comparing your treatment to that of fellow chargers. It is about whether you think the (choices for) parameters and weights are fair. (D6)

Fairness is also invoked on the level of messaging. One of the aims of the design project is to convey a "positive message" (D4) about the municipality's role in the transition to fully electric driving in the city.

> The core of the message is that the interests of different parties are fairly represented in order to arrive at solutions that work for citizens, government, and private parties as smoothly as possible. (D4)

Project members agreed this message should be conveyed using a "positive tone of voice" (D4). When discussing tone of voice, fairness is once again invoked, although it may also be understood as truthfulness, because the Dutch word for both truthful and fair is the same ("eerlijk"), and it is not entirely clear from context which meaning is intended here.

> The design and all communication around it are based on a positive tone of voice (*truthful*, predictable, not too difficult, positive connotation, municipality listens, no algorithmic doom scenarios). (D4)

We see truthfulness and fairness recur on different levels throughout the project. Truthfulness is seen as a quality of information provided by the system, producing transparency about how a person is treated. This treatment can be

more or less fair, the assessment of which is enabled by truthful information. The system is imagined to convey a message that people are indeed being treated fairly. That every party with an interest in smart EV charging, citizens included, is given fair consideration. Finally, fairness and truthfulness are (ambiguously) invoked as a desired *tone* of messages.

2.5.2 Citizen Experience of a Transparent AI System

Theme C1: Transparency Mediates Concern

In general, participants were welcoming of automated decision-making in the EV charging process. Many responded positively to the notion of using automation to optimize EV charging towards what could be described as common interests: a stable electric grid, a fair distribution of power, and sustainability in general terms.

Also, many people seemed more or less indifferent to the presence of automated decision-making. For example, when asked if any automated decisions had been made, P6 responded "Yes, but based on what was already there." By which they meant the system was simply responding to the inputs it sensed in the environment. P5 commented they were sure there were "technicians who have thought about it ..." In other words, they put their faith in the expertise of the people who built the system. P6 simply stated "I take it the way it is."

One of the most striking statements for us was when P7 said "I don't think I should be able to make a choice about that," referring to trade-offs between collective interest at the expense of individual efficiency.

People's indifference to AI may be in part due to the fact that when charging in the city, less is at stake compared to, say, a fast charging session. Charge speed is slow, session duration is short, and out of all charging options (at home, at work, at a fast charging facility, in the city), public charge points are the least depended on. P7: "on the one hand, when I'm going to run an errand, and I'm done within the hour, yeah, then I don't care how fast." Any charge received while parking is considered a bonus.

A few participants *did* express concerns about situations in which they would be disadvantaged by the system and the impossibility of making a one-time exception. For example, when they were in a hurry or when they were forced to charge during peak hours.

The strongest reactions *against* automated decision-making related to the shared car priority feature. Many participants latched on to this while ignoring most of the other rules made transparent. P2: "Shared car has priority. I

don't like that, but okay. Sustainable, of course." Some recognized it would be beneficial for sustainability reasons, so they did accept its rationale. However, none of the participants were shared car drivers themselves. Some participants wondered about what was considered a shared car, who determined this, and how the system would deal with, for example, shared cars from outside of Amsterdam.

P2, a resident of Amsterdam, made a connection between shared car priority and local politics, which recently had taken a more left-leaning, progressive turn than in years before. They expressed fear of politicians pushing for more extreme forms of shared car priority at the expense of private car owners.

> Well ... See if it will be that way later ... You will get a political decision. Politicians are going to say yes, but ... If there is a shared car, the other cannot get in, and so on, you have to let three shared cars go first. Especially with Femke[5] in Amsterdam, I am a little afraid of that. (P2)

The comment was made somewhat in jest, but it does stand in remarkable contrast to the general indifference to automated decision-making that we have tried to capture thus far.

This discomfort with treating some EVs differently from others may be due in part to the scarcity of charge points. It can be a challenge to find a free spot. If one ends up next to a shared car and charges slower as a result, it feels unfair.

> Well, that shared car [priority] makes me go, gosh darn it ... I find it very annoying. At a busy time, I was racing through the city, and all the stations in the neighborhood were occupied. Then I arrive here, and then I actually get punished a bit more. Then they would have had to put a few more stations in Zuid [an affluent area of the city] ... We all have a Tesla. So that's a bit complicated. (P2)

There appears to be a relationship between people's attitudes towards automated decision-making and the purposes to which it is put. This is different from the narrative about people being suspicious of all automation, regardless of where it is applied. It also sheds a different light on in which cases transparency is necessary or desired.

Theme C2: Transparency Is Burdensome

With this theme, we want to capture how EV charging is often a sub-optimal experience, made worse by the additional demands transparency puts on people.

5. Femke Halsema, at the time mayor of Amsterdam and former leader of the national Green Left party.

First of all, EV charging, in general, is an error-prone activity. Poor design and engineering of charge points and wider infrastructure frequently lead to failed charging attempts. In our prototype evaluations, most participants started a charge session in the "wrong" way, even though instructions were listed on the opening screen. They typically made their way through it in a trial-and-error fashion: swiping a card and plugging in a connector in succession until the system progressed to the next state, not taking time to read any instructions beforehand. It is a usability truism users do not read, and people charging an EV are clearly no exception. P7 acknowledged as much when she responded to the explanation by saying "So anyway, that only makes sense if you read it very carefully." It is likely this situation is *even worse* outside of a prototype evaluation because when EV charging at a public charge point, people are likely to be in a rush. They might have someone else waiting to use the same charge point, and in any case, they will probably have somewhere else to be. So, as P7 pointed out, they are not inclined to study a user interface at length when they are setting up their EV for charging.

> In the city, I am not going to do that … I think. Certainly not when I go shopping … usually it is like, let's get it over with, and then you want to go on again … I would be very interested in how it works, but I would rather see that afterward. (P7)

A final source of unease is uncertainty over the amount of charge delivered. While charging, prototype v1 displayed the real-time charge speed in amperes (amp), a measure of current. This was a largely meaningless indicator for participants.

> This—12 ampere doesn't tell me very much. I really benefit from seeing where I am at now and how much time it will take me to get to 100%. So what percentage am I at, and how much time does it take me to get to 100%? I think that's important. (P5)

V2 switched to kilowatt (kW), a measure of power. Participants could at least extrapolate from this real-time measure to an expected amount of energy received at session end. Most participants were also able to translate a session's worth of charged energy measured in kWh to range because they had memorized the capacity of their EV battery. Or, they compared the listed amount to what they knew a fast charge point delivers. Needless to say, all this mental arithmetic meant more work for participants, and although some did take pride in their ability to perform it, most were perfectly happy to offload all of it onto a system.

For all of these reasons, it should come as no surprise many participants reported feeling overwhelmed by explanations. P9: "There is already a lot of information on it, I must say." Participants do not welcome additional demands

put upon them by this information when all they want to do is charge their EV. P3: "I think it's a lot of information. … I just want to charge." Additional information lead to confusion. P7: "I think it's too much info. Honestly, it's confusing. From the start, I actually think I see way too much." This confusion is caused, at least in part, because participants think they are expected to act on it somehow.

Apparently, participants are focused on completing the task of EV charging with confidence and minimal hassle. The information added to the interface in the interest of transparency does not directly support task completion. Because, as we captured with the previous theme, participants are largely indifferent to AI, this information is experienced mostly as a burden.

> This is kind of competing for my attention. … What I want to know is just steps 1, 2, 3, and 4 of my actions, but this here is a lot of information that makes me go "what should I do with it?" (P9)

Theme C3: Transparency Invites Strategic Behavior

Participants expressed intent to adapt their behavior to the system, something that could be considered an unintended side-effect of providing more transparency. Participants suggested they might pick a different time to charge so that they would benefit from increased speed during off-peak hours or would enjoy extra speed when the sun was out. Another reason for changing behavior, particularly in relation to solar power availability, was sustainability.

Some participants were driven less by a desire to be more sustainable and were more interested in reaping economic benefits. For example, P5, while discussing the "receipt" displayed at the end of a charging session, talked about how they were most interested in changing their behavior so they would *pay less*.

> What I would like to see here is … What have I paid? … What can I do better to charge better next time? Now I'm looking at this, but I don't immediately see what I can do about it. Do I have to go to the green bar? How can I influence it? (P5)

Regardless of whether a participant was looking to improve charging speed, sustainability, or cost, the availability of more information about smart charging system operation appears to inspire an intent to optimize behavior.

It should be noted not all participants were as keen on changing their behavior. Most participants happily speculate about what *other* people might do with the information provided. P8: "Well, then people are aware, and they charge at a certain time when they have the choice." But when we put them on the spot, they frequently admitted they did not expect to change anything about their behavior *themselves*. This could be because benefits do not outweigh additional effort. It

could also be because, particularly in the context of city charging, the time and place of charging sessions are strongly dictated by circumstances, such as the availability of free parking spots. As both P1 and P8 stated: "if I have to charge, I have to charge."

Theme C4: Transparency Evokes Desire for Control

Many participants interpreted explanations *themselves* as things to be interacted with. When confronted with the opening screen, P1 asked "But do I have to choose something or not?" and while looking at the status indicators said: "What those are? … A choice, I can tap on" (P1). And P3 commented "I think that's very interesting for you. That people look at such a screen in this way. They are staring at it going 'what the hell should I choose.'"

Most participants expected to be able to change things to their advantage, not only in terms of charge speed but also in terms of price: "Here I can say 'shared car yes or no'" (P4). "What I see here I could determine myself … Determine the best price-quality ratio" (P5). "That you can interact. So if I do that, it will be cheaper or more expensive. Or it will go faster or slower. You kind of think that" (P2).

In some cases, participants wanted to have the choice to be altruistic. P1 expressed a desire to decide for themselves if they would indeed give priority to the shared car connected to the charge point they were using: "I gave something to the shared car. So I've done my good deed for the day." It should be noted being nice to shared cars would really depend on the circumstances. "Shared car priority. So here you can choose how nice you want to be. But maybe also if you are in a hurry, then you are in a hurry" (P1).

Those who were more or less indifferent to AI typically did not respond too strongly to the revelation they, in fact, could not exercise any control. Some participants, however, did respond with some chagrin, such as P3 here:

> Well, it is clear to me that I actually decide almost nothing, and the system decides everything for me. I find that strange… Because I… Well, the system itself makes decisions, and I don't know what exactly is happening. (P3)

P3 clearly did not experience the sense of agency expected to result from transparency. Other participants warned against not including user choice because that would lead to rejection from users: "I think that people will respond negatively more quickly maybe … Because then you can only grumble about it" (P1).

Furthermore, lack of control in some cases leads to participants questioning the value of including any explanations at all.

> Just now, I had the idea that I could opt for green energy. Or I can choose to deliver back to the net. So, I thought I could make choices. But basically, it's just you plug in, and all kinds of stuff happens. But I have no influence on it. So then I think, why should I have all that information? *If I have no influence on it, what should I do with it?* (P3)

When asked how they would deal with automated decisions they disagreed with, the majority of participants seemed somewhat resigned to accepting the system's functioning as it was presented to them. Although one might expect some participants would want to make their disagreement known to system operators, most stated they would simply try to find a different (non-smart) charge point.

> If I know I have no influence on things right now because I can't change any settings when I know that a decision is made for me and I do not like them at that time, or faster loading or yeah, that car is charging really fast because it is a shared car, well, in that case, I can choose to find another charge point. (P4)

Making attempts to influence system developers hardly ever came up. P1, when asked what they would do if they disagreed with how things worked, said: "Well, I tweet. I'll put it on Twitter." This same participant felt if you cannot make any choices, if everything is automated, this is bad, because "then you can only grumble about it."

2.6 Discussion

Using a reflexive thematic analysis of design process documents and prototype evaluation transcripts, we have captured ways in which a group of experts understand transparency and how the transparent urban AI system resulting from their efforts, a transparent smart EV charge point, is experienced by citizens. In this next section, we reflect on three tensions constructed from a comparison between the two groups of narrative themes (Figure 2.4). These are: (1) information quality over quantity; (2) level of concern; and (3) sense of control.

2.6.1 Tension #1: Information Quality Over Quantity

We have found that according to experts, transparency is created by providing truthful *information about automated decisions (X1).* However, the belief that being truthful leads to increased trust is not born out by previous research. For example, placebo explanations can still improve trust [94]. Also, explanations can be satisfying to users without necessarily being truthful [97]. What is more, users

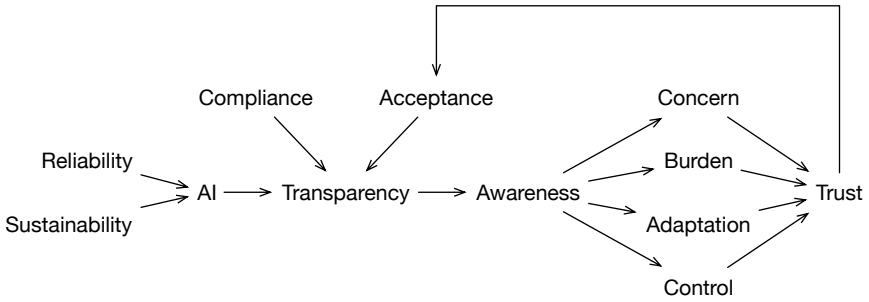


Figure 2.4
Diagram illustrating tensions between expert motivations and citizen experiences. Experts implement AI to achieve energy infrastructure reliability and sustainability. A need for legal compliance and societal acceptance drive efforts to make AI transparent. This transparency increases citizen awareness, which triggers several responses: concern about optimization targets, increased cognitive burden, various forms of adaptive behavior, and an increased desire for control. Each response impacts citizen trust in positive or negative ways, which in turn will affect social acceptance desired by experts.

typically cannot ascertain *correctness* of system output from an explanation alone [284].

Experts believe that, because automated decisions might benefit some more than others, and because AI is by its nature hidden, they need to be made visible *(X1).* This can be seen as an example of setting too high standards of explanation for machine decisions [369]. A vast number of decisions are made by city governments, benefiting some more than others, and a lot of city governance recedes from the view of ordinary citizens. Not all such decisions are made visible to the extent pursued in this project. On the other hand, it can be argued AI systems demand a higher degree of transparency precisely *because* of their technical nature. When ML techniques produce proper black-box models, such systems are fundamentally less predictable and more opaque than a human equivalent [139].

The project can be considered an example of how pursuing transparency sidelines the question "should we be using AI at all?" [71]. Experts did not question the decision to use an AI system in EV charging. The pursuit of transparency serves to support the ongoing use of AI.

Experts in our study talk about transparency in terms of *making the hidden visible*, as if an AI is something that can be seen and there is merely something obstructing our view in need of removal. However, what we see them *do* in our project is something quite different and supports the notion that transparency requires active production of information [237]. The language our experts use

is also at odds with the view that understanding a system requires more than seeing inside it, that it requires being able to change systems and seeing how they behave in relation to their environment [11].

Information provided in the interest of transparency is frequently experienced by citizens as burdensome (C2). It is *not* perceived to be supportive of charging an EV, a task rather error-prone and stress-inducing to boot.

Citizens apparently perceive the benefits of engaging with explanations do not outweigh the costs. This can be due to explanations lacking completeness [200], although the interface in question did provide explanations of a range of system aspects. A more likely reason is that citizens believe they can form an adequate mental model of the system simply by *using* it, where "adequate" may be next to no model at all [50].

In other words, information *quality* over quantity is the actual moral problem of transparency [237]. The information produced in the interest of transparency can occlude as much as it reveals and, in the process, add to anxiety already felt in response to being subjected to automated decision-making [11, 53]. Others have argued for a distinction between explanations in direct support of tasks and explanations of AI decisions with only an indirect connection to user actions [284]. Our project more closely matches the latter category. Perhaps, when AI is indirectly connected to user actions, explanations should be made subordinate to information that is in support of tasks and made available upon explicit user request or when a high likelihood of a need for explanations has been detected through implicit signals.

There exists a tension here between experts' desire to make citizens aware of the presence of an AI and the aim of old-fashioned user interface design in support of a user's task. In a world where AI is part of an increasing number of systems, we should ask ourselves if user interfaces are the proper location for raising awareness.

This point of tension is of particular importance when dealing with street-level touchpoints of urban AI systems because people's attention tends to be even more limited due to the pressures of everyday urban activities.

2.6.2 Tension #2: Level of Concern

Experts pursue transparency because citizens may reject the use of automated decision-making in public infrastructure (X1). The position of experts here mirrors the idea that transparency has no moral content [237]. It is a means towards an end. For our experts, the goal is acceptance by the public, EV drivers in particular. Various factors may lead to this acceptance, but the reasoning in the project seems to

be that once citizens understand the workings of a system, they can ascertain its fairness, and once they know it is fair, they will accept the system. However, the literature we have reviewed provides a mixed view of the relationships between transparency, understanding, fairness, and trust [28, 60, 188]. What this means is that if acceptance is the aim, relying on transparency is likely insufficient or may even backfire, for example, when an excess of information decreases people's trust [188]. It may also be the case that a limited form of transparency offering justification of decisions is more likely to increase the perceived legitimacy of a system, rather than the more extensive transparency of the automated decision-making process *itself* attempted in this project [106].

If we consider the responses of some citizens to the explanations provided, it appears trust may have *decreased* instead of improved. Transparency enabled them to see some of the decisions affecting charge speed and, in some cases, had them wondering about system developers' motivations to include those factors, most notably in the case of the shared car priority feature. Being critical is not necessarily the same as being distrustful, but uncritical acceptance is unlikely to be the result of transparency. This suggests transparency efforts should be prepared for public debate with users around issues uncovered by transparency. Furthermore, such opportunities for "voice" may, in fact, *increase* trust [154].

The experience of citizens is characterized by an overall acceptance of, or even indifference to, the presence of automated decision-making (C1). AI is seen as a convenient way of optimizing for broadly shared collective interests such as electric grid stability and sustainability.

This echoes others' findings that people find AI useful, but crucially also that the AI systems people find *most* useful are not necessarily the ones they find most fair [78]. Potentially, also in our project, people are making trade-offs between fairness and usefulness. This could go some way towards explaining our participants' general indifference to the use of AI in EV charging.

Only when something is at stake (illustrated in our study by the shared car priority feature) do citizens start to question AI (C1). Others have pointed out different people respond to transparency in different ways. As a result, people also *trust* a given system in different ways [11]. The various motivations our participants have for EV driving (e.g., saving money or protecting the environment) may influence how they respond to AI features exposed to them in the interface. Prior experiences with organizations deploying AI influence the extent to which people trust AI [48]. Similarly, the extent to which participants consider local government, power companies, etc. to be "on their side" also figures into how they perceive the use of AI for automated decisions.

AI opacity *as such* is hardly ever an issue for citizens. This suggests experts should focus more on contested issues (such as air quality, parking space, congestion) and how automated decision-making *interacts* with those. A typical line of reasoning is that technology can improve these issues and should, therefore, be welcomed. This is certainly a general driver behind the push for electric mobility in Amsterdam. At the same time, it is felt AI lacks transparency and that this should be fixed. What is *not* considered is how a person's view on an issue like spatial justice may affect the extent to which they welcome EV charging and, by extension, a smart charging system, *regardless* of how transparent it is. This suggests transparency efforts should be focused more on those matters that are actually contested and how AI *mediates* those issues.

2.6.3 Tension #3: Sense of Control

Experts believe explanations are actionable *by citizens (X2).* Experts presume explanations make it possible for citizens to assess the *fairness* of decisions by evaluating inputs, processes, and outcomes of "the AI," by having access to a justification for the AI's design and by knowing who "owns" the AI.

However, actionability is influenced not only by content but also by the format of explanations. For an explanation to be actionable, it must be usable as "currency" in a person's decision-making process [344]. Having the ability to assess fairness by itself does not equip a person to *act* on that assessment. The current transparent EV charge point design addresses this issue through the previously described "receipt" feature (Section 2.4.2). However, none of our participants spontaneously suggested they would use those resources if they disagreed with the system's functioning.

Citizens intend to use explanations as a resource for adapting behavior towards altruistic or egoistic ends (C3). The tension between transparency and the possibility of "gaming" behavior, as well as concerns over the exposure of intellectual property, has been noted previously [10, 173]. In our project, we did not see experts express such worries.

The fact that citizens indicate a desire to change behavior in response to explanations suggests information *is* actionable to some degree. However, actionable information by itself does not appear to provide sufficient means of influence over system behavior.

Explanations created expectations of user control, an ability to override automated decisions (C4). The absence of control leads some participants to question the relevance of explanations.

Others have shown transparency paired with *control* can alleviate anxiety caused by automated decisions [173]. Users can have favorable cost-benefit perceptions of explanations if they are able to act on provided information by adjusting system behavior [200]. Possibly, our participants' responses to the transparency interface would have been quite different had it also offered means of directly or indirectly influencing the operation of the AI. This underscores the fact that explanations by themselves are not always perceived as sufficiently actionable. Furthermore, it is also possible that, rather than alleviating frustrations around lack of control through something akin to seamful design [10, 59], further opening up of systems actually produces even greater anxiety in users who are already overwhelmed by explanations.

This desire for control appears to be at odds with the fact that experts are not willing or able to offer direct control over system behavior and anticipate explanations alone to be a sufficient form of accountability. Citizens do not want to or believe they are not able to, petition experts for changes to system behavior despite the presence of explanations that could be leveraged for purposes of recourse. This suggests more explicit channels for voice should be made available in or around touchpoints of AI systems.

In case of disagreements with automated decisions, most citizens opt to defect to an alternative means of charging rather than try and influence policies shaping system behavior (C4).

Some have argued transparency invokes a neoliberal model of agency [11, 56, 186]. We think our project is a clear illustration of this logic in action. Our participants did not appear to feel they had a substantial say in the operation of the system. This could be because there were no clear "channels for voice". It could also be because people's lack of "loyalty" to the organization deploying the system made them disinclined to go through the trouble of acting on their disagreements [58, 154]. In any case, we feel that in our project, transparency alone was insufficient for creating any form of agency beyond EV drivers simply not "buying" the services of the charge point they disagree with, a reduction of citizen participation to consumer choice. The implication of this for experts is that if shared control over system behavior is desired, "voice" needs to be put on the table. Citizens need to be able to discuss and debate the significance of the information they are provided through transparency in a dialog with system designers, developers, and operators. This suggests efforts to make urban AI more transparent should be paired with more participatory and collaborative approaches to city governance and policy-making [82, 111, 116].

In closing, we argue these tensions show transparency should not be seen as a property of technology but must be understood as a communicative process between experts and citizens, who are more than mere users. AI systems mediate this process, inviting some actions and inhibiting others [348]. Understanding a system is not the product of simply receiving and processing information. Understanding emerges from debates between stakeholders and is always provisional. All three tensions we identify (information quantity, level of concern, desire for control) in various ways point to the need for additional channels for voice through which this debate can be facilitated.

Some of our findings relate quite specifically to the vagaries of smart EV charging. An example would be the challenge of finding an intuitive measure of charge speed. Furthermore, the class of AI we worked on is a deterministic one. The design solutions pursued in our project may not be viable when dealing with stochastic systems. At the moment, EV driving is something accessible mostly to significantly affluent or professionally employed people. Certainly, our sample of citizens skews highly educated. It is likely this informs their attitudes to issues such as sustainability and automation. Our expert-citizen distinction leaves out a lot of other relevant stakeholders to investigate. Indeed, our citizens were all direct stakeholders and users of the system. There are plenty of citizens who are not EV drivers but who are likely to be impacted in various ways by the roll-out of EV charging infrastructure, e.g., by reduced neighborhood parking space or by continued prioritization of road space for cars. There are also plenty of stakeholders who have relevant expertise but who are not considered "experts" in the sense we have been using here. That is to say, those stakeholders who have some level of formal influence over the shape the system takes.

2.7 Conclusion

We have presented findings from a practice-based design research study investigating diverging conceptions of transparency by expert and citizen stakeholders of an urban AI system. Our expert participants believe transparency is achieved by providing truthful information about automated decisions. They expect citizens to be able to assess system fairness using this information and be able to act on this information. Meanwhile, our citizen participants are largely indifferent to AI; they primarily experience explanations as burdensome and question their relevance if they are not accompanied by the ability to override system decisions.

Transparency is a growing topic of interest in HCI design research, and in public discourse, it is commonly invoked as a solution to the negative effects of AI opacity. As a result, transparency has also been taken up as a desirable

system property in urban AI systems development. Our findings illustrate that it is necessary to remain critical of assumptions driving the pursuit of transparency in AI system user interfaces. Transparency puts additional cognitive demands on people and shifts the responsibility of ensuring fairness onto them, reinforcing a neoliberal model of agency.

For these reasons, we believe transparency should be reframed. It should not be seen as a property of a system through which information flows from experts to individual users. Rather, transparency must be seen as a communicative process between experts *and* citizens, mediated by AI systems.

Acknowledgments

We want to thank our field study participants for their rich responses and the UI for Smart EV Charging project consortium members for supporting our participation: the Amsterdam Institute for Advanced Metropolitan Solutions, ElaadNL, The Incredible Machine, and the municipality of Amsterdam.


HARCHAVING

Chapter 3

Contestable AI by Design: Towards a Framework

Citation:
Alfrink, K., Keller, I., Kortuem, G., and Doorn, N. "Contestable AI by Design: Towards a Framework." In: *Minds and Machines* 33.4 (Aug. 2022), pp. 613–639. DOI: 10/gqnjcs

Abstract:
As the use of AI systems continues to increase, so do concerns over its negative social consequences. 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 of the 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 design frameworks. In this article, we use qualitative-interpretative methods and visual mapping techniques to extract sociotechnical features and practices from the literature that contribute to contestable AI and synthesize these into a design framework.

3.1 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, including the undermining of the democratic rule of law and the infringement of basic human rights to dignity and self-determination [e.g. 61, 67]. A way to counteract such harmful automated decision-making is through *contestability*. Contestable AI systems are open and responsive to human inter-

vention 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 [9, 146, 153, 223, 298, 341, 343]. 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" [161, 214] 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 the 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 typical AI systems 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.

3.2 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 [326]: "[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 the contestability of AI systems in general and in automated decision-making, specifically [9, 153, 298, 341].

Hirsch et al. [153] 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 who are 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. [341] argue that contestability can surface values, align design practice with the context of use, and increase the perceived legitimacy of AI systems. Contestability is a "deep system property" representing a coming together of humans and machines 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 [9] 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 [177]. 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 [149]. 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 [298], 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 one's 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 that was 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 subjects and human controllers. 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").

3.3 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 [114, 197]. What is more, AI systems are distinct from "traditional" sociotechnical systems because they include "artificial agents" and humans interacting in a dynamic evolving environment [280]. 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 [83, 180, 334]. 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 [80, 264].

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 into their own practice. To illustrate this point, we briefly summarize a number of prominent systematic reviews and meta-analyses drawn from across disciplines [174, 247, 310].

Jobin et al. [174] 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. [247] 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 [310] 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 and 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 the 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.

3.4 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" [214]. *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 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 [321], 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 Obrenovic [268]. 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.

3.5 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 a number of elements (features and practices) contributing to contestable AI. Finally, we used visual mapping techniques to synthesize these elements into a pair of framework diagrams.

3.5.1 Data Collection

Our data-collection procedure broadly follows Moher et al. [244]. 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 terms for AI. See Table 3.1 for an overview of search terms used. The exact Scopus search is as follows:

```
TITLE-ABS-KEY((design*) AND (contest* OR controvers* OR
debat* OR disagree* OR disput* OR dissen*) AND
("artificial intelligence" OR "AI" OR "machine learning"
OR "ML" OR algorithm* OR "automated decision making")) AND
(PUBYEAR > 2015) AND (PUBYEAR < 2022) AND
DOCTYPE("cp" OR "ar")
```

We collated the results and first removed duplicates. Then, using Rayyan [271], 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 of the new inclusions. Figure 3.1 shows the stages of our systematic review.

Table 3.1
Search terms used.

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

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

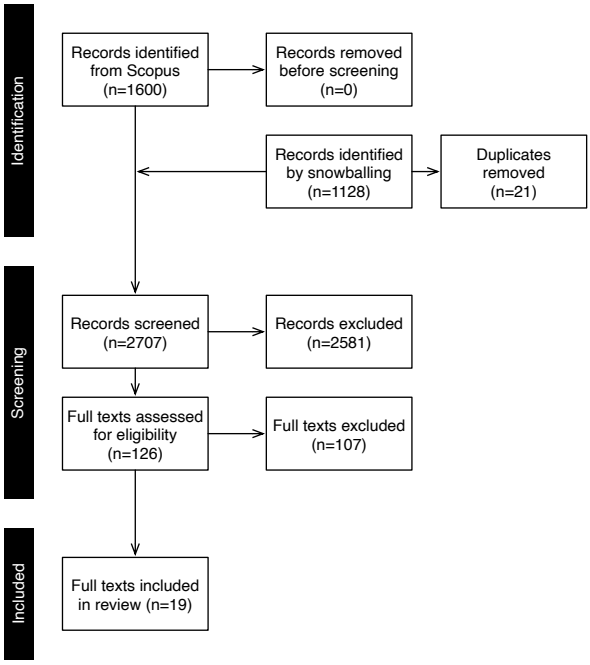


Figure 3.1
Flow of information through the different phases of the systematic review.

3.5.2 Analysis and Synthesis

Our approach to analysis and synthesis is adapted from reflexive thematic analysis as described by Braun and Clarke [45]. 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.

3.6 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 Table A.1, Table A.2, and Table A.3 and are described in detail in the following sections.

3.6.1 Features

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 a disagreement between both systems occurs, the decisions 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 [9, 95, 223].

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 [4].

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 [190, 266, in 341]. 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 [22, 153, 341, 342]. 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 [153, 172].

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 [223]. Explanations should contain the information necessary for a decision subject to exercise their rights to human intervention and contestation [22, 223, 278].

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 [4]. 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 [22]. The norms decisions should adhere to should be elicited and specified ex-ante [4].

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 [4, 9, 47, 66, 153]. This behavioral approach accounts for the limitations of transparency efforts focusing on "the algorithm" in isolation [11, in 146]. It also seeks to strike a balance between usability and comprehensiveness in an effort to avoid the "transparency paradox" [262, in 66].

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) [22].

Although explanations may be of a static form, if deep understanding and exploration of counterfactual scenarios are desired, "sandboxing" or "black box in a glass box" approaches are worth considering. These approaches enable users 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 [162, in 153]. By offering

explanations that include confidence levels, human controllers can direct their focus to those decisions warranting closer scrutiny [153, 341].

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 [203, 226, 288, 333, in 91, 153].

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 [291, in 146].

Human Review and Intervention Requests

This feature is aimed at decision subjects and third parties acting on their behalf. It gives subjects the ability to "ask questions and record disagreements," both on the individual and the aggregate scale [153, 278, 341].

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 [189, in 146].

A commonly recommended mechanism for responding to post-hoc contestation is human review and intervention [223]. 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 [9]. Such intervention requests could be facilitated through auxiliary platforms or be part of the system itself [9, 22]. 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 [223].

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 [9, 354].

In the context of GDPR, the right to human intervention is tied to fully automated decision-making only [47]. 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 [9]. By extension, human controllers who respond to intervention requests should have the authority and capability to actually *change* previous decisions [47].

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 [9]. AI system controllers must make room within their organizations for receiving, evaluating, and responding to disputes [298].

Channels for contestation should be clear, accessible, affordable, and efficient so that further harm to subjects is minimized [223, 343]. Mechanisms for requesting human intervention should provide "scaffolding for learning" [14, 296, in 342]. Documentation of the decision-making procedures should be integrated with the appeal procedure and communicated in alternative formats to ease comprehension [342] and to help subjects in formulating their argument [223, 343]

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 [343] or having third parties represent subjects [22, 91, 223, 342].

Another limitation of current appeal procedures is that they handle decisions individually [341]. Groups should be able to acquire explanations of decisions collectively. Developers should not only consider individual impacts but also group impacts [91]. Mechanisms for contestability should allow for collective action because harms can be connected to group membership [223]. 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 [231, 260, 297, in 342]. Subjects should be supported in connecting to those who share their fate [343].

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 [22]. 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 [236, in 47]. Therefore, contestation includes human intervention but should not be *reduced* to it. Care should also be taken to prevent contestability from becoming merely

a way for subjects to complain about their plight. This means contestations of this kind cannot be handled in a fully automated fashion because a dialectic exchange between humans and machines is not possible in a meaningful sense. 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 [298]. Contestability should include a right to a new decision, compensation of harm inflicted, or reversal [223].

Tools for Scrutiny by Subjects or Third Parties

This feature supports scrutiny of AI systems by outside actors (decision subjects, indirect stakeholders, third parties), 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* [223]. 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 implementation. Compliance is established by tracing this link through requirements, specifications, and the code itself.

Documentation should describe the technical composition of the system [342]. 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) [305, in 9], the role of human decision-makers, group or systemic impacts and how they are safeguarded against [223]. Mitchell et al. [241] and Gebru et al. [121] 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. [146].

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 [65, in 9].

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 [199, in 9].

3.6.2 Practices

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 [47, 146, 298], and pay close attention to how the system may "mediate" new and existing social practices [347, in 153]. 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)) [91, 223]. Such assessments can also enforce the production of extensive technical documentation in service of transparency and, by extension, contestability [22]. 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 [303, in 179]. If system decisions are found to impact individuals or groups to a significant extent, contestability should be made a requirement [146]. 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 [9, 47, 354]. 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 [223].

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 [22, 91].

Agonistic Approaches to ML Development

This practice relates to the early lifecycle phases of an AI system: business and use-case development, design, and procurement of training and test data. 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" [149, in 9]. An agonistic approach to ML development allows for decision subjects, third parties, and indirect stakeholders to "co-construct the decision-making process" [341]. The choices of values embedded in systems should be subject to broad debate facilitated by elicitation of the potentially conflicting norms at stake [146]. 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 [9, 91]. In HCI, a well-established means of involving stakeholders in the development of technological systems is participatory design [69, in 9]. By getting people involved in the early stages of the AI lifecycle, potential issues can be flagged before they manifest themselves through harmful actions [9]. 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 [9]. Representation can be very direct (similar to "jury duty"). Or more indirect (volunteer or elected representatives forming a board or focus group) [343].

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 [127, 175, in 179].

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 [179].[2] AI systems, like all designed artifacts, embody particular political values [360, in 66]. A participatory, agonistic approach should be aimed at laying bare these values and creating an arena in which design choices supporting one value over another can be debated and resolved (although such resolutions should always be considered provisional and subject to change) [179]. König and Wenzelburger [192] 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. [313].

Quality Assurance During Development

This practice ensures safe system performance during the development phases of the AI system lifecycle. This includes the collection of data and training of models, programming, and testing before deployment. A tried and true approach is to ensure the various stakeholders' rights, values, and interests guide development decisions. Contestability should not be an afterthought; a "patch" that is 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 [189, 190, 207, in 9, 153, 179]. Before deploying a system, it can be tested, e.g., for potential bias, by applying the model to datasets with relevant differences [278]. 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 [179]. 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 [153, 341, 342]. Developers should seek feedback from stakeholders both with respect to system accuracy and ethical dimensions (e.g., fairness, justice) [354].

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 of "procedural regularity" captures the idea that one should be able to determine if a system actually does what it is declared to do by its developers. In particular, when models cannot be simplified, additional measures are required to demonstrate procedural regularity, including monitoring [22]. System operators should continuously monitor system performance for unfair outcomes both for individuals and, in the aggregate, for communities. To this end, mathematical models can be used to determine if a given model is biased against individuals or groups [131, in 9]. 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 [153, 342]. 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 [354].

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 are 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 [153, 223, 278, 341, 342].

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 [22, 91, 95, 341]. System operators may be obligated to implement model-centric tools for ongoing auditing of systems' overall compliance with rules and regulations [22]. Companies may resist opening up proprietary data and models for fear of losing their competitive edge and users "gaming the system" [66]. 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 [22]. Such efforts can be complemented with the use of technological solutions, including secure environments that function as depositories for proprietary or sensitive data and models [91].

3.6.3 Contestable AI by Design: Towards a Framework

We have mapped the identified features in relation to the main actors mentioned in the literature (Figure 3.2): **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.

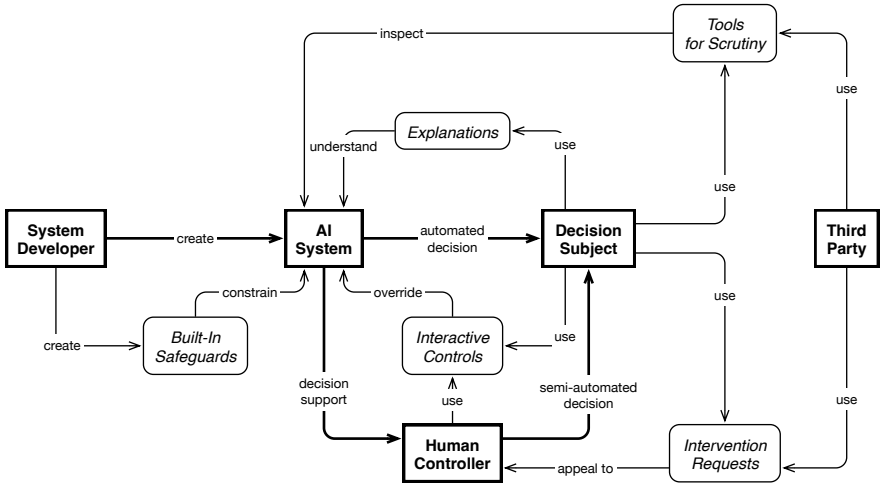


Figure 3.2
Features contributing to contestable AI.

We have mapped the identified practices to the AI lifecycle phases of the Information Commissioner's Office (ICO)'s auditing framework [29] (Figure 3.3). These practices are primarily performed by system developers. During **business and use-case development**, *ex-ante safeguards* are put in place to protect against potential harm. 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.

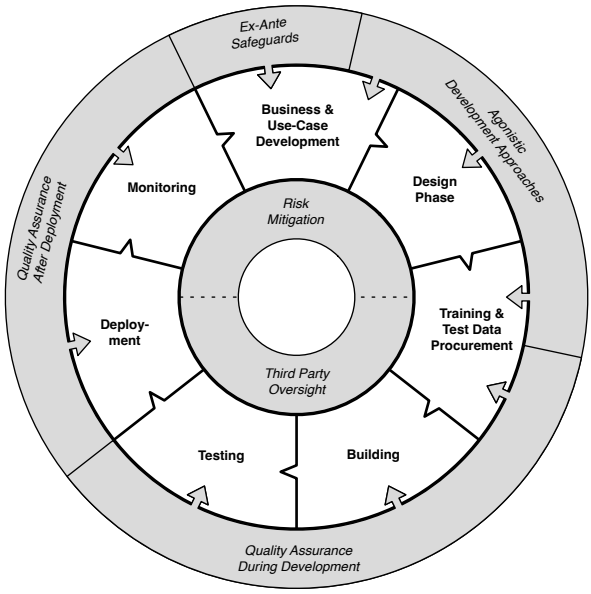


Figure 3.3
Practices contributing to contestable AI.

3.7 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 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; *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 included papers 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 do 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. [247] 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. 63, 165, 245, 285].

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 the 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. 38, 57, 238, 312, 339, 349, 362].

Much of our own empirical work is situated in (local) government public services in OECD countries. Some distinctive features of such settings include the 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 upfront 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 instead center democratic values of inclusion, plurality, and justice.

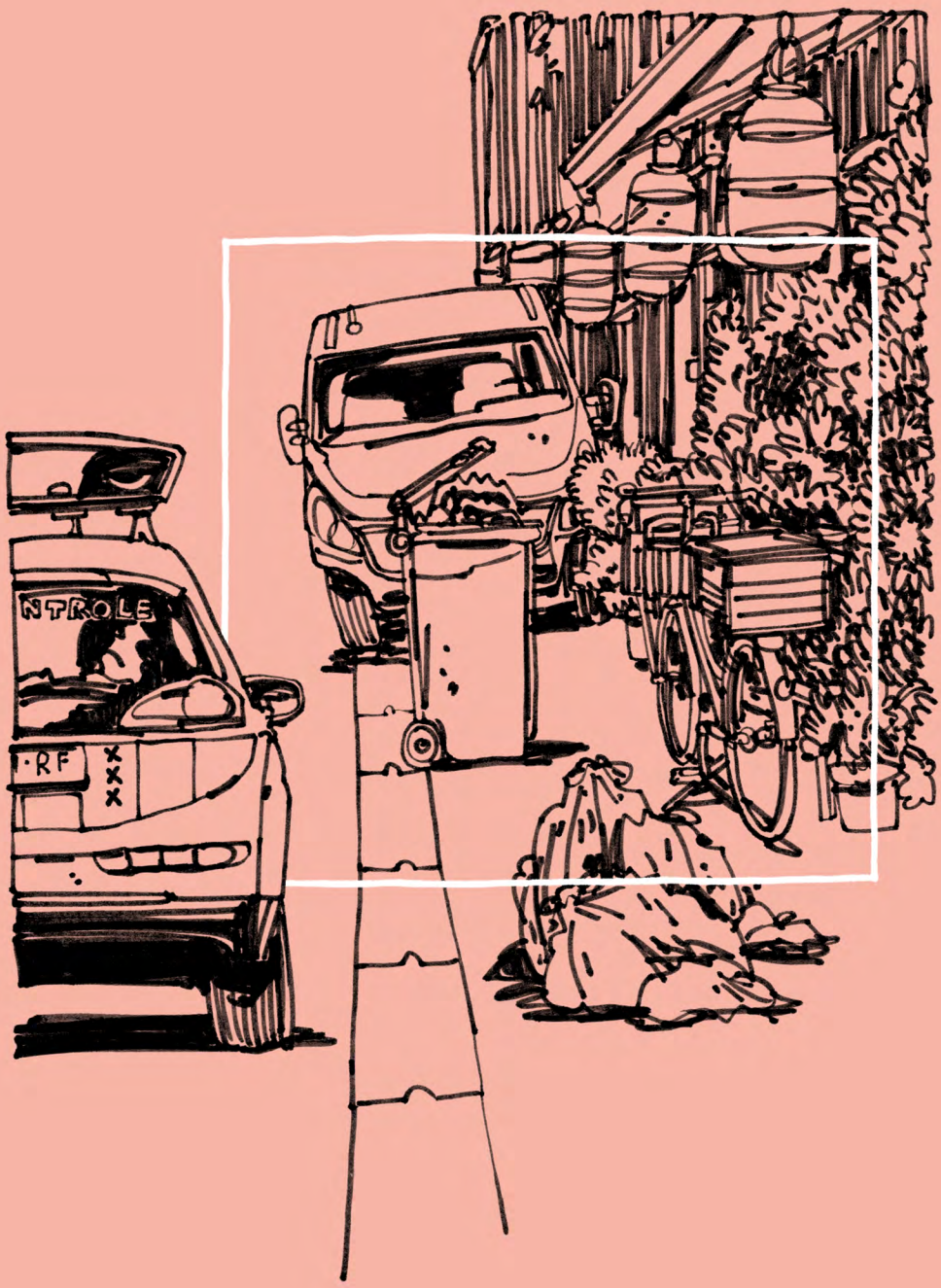
3.8 Concluding Remarks

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.

Acknowledgments

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


NTROLE
RF

Chapter 4

Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute

Citation:
Alfrink, K., Keller, I., Doorn, N., and Kortuem, G. "Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, pp. 1–16. DOI: 10/gr5wcx

Abstract:
Local governments increasingly use artificial intelligence (AI) for automated decision-making. Contestability, making systems responsive to dispute, is a way to ensure they respect human rights to autonomy and dignity. We investigate the design of public urban AI systems for contestability through the example of camera cars: human-driven vehicles equipped with image sensors. Applying a provisional framework for contestable AI, we use speculative design to create a concept video of a contestable camera car. Using this concept video, we then conduct semi-structured interviews with 17 civil servants who work with AI employed by a large northwestern European city. The resulting data is analyzed using reflexive thematic analysis to identify the main challenges facing the implementation of contestability in public AI. We describe how civic participation faces issues of representation, how public AI systems should integrate with existing democratic practices, and how cities must expand capacities for responsible AI development and operation.

4.1 Introduction

Local governments increasingly use artificial intelligence (AI) to support or entirely automate public service decision-making. We define AI broadly, following Suchman [326]: "[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." As the use of AI in public sector decision-making increases, so do concerns over its harmful social consequences, including the undermining of the democratic rule of law and the infringement of fundamental human rights to dignity and self-determination [e.g. 61, 67]. Increasing systems' *contestability* is a way to counteract such harms. Contestable AI is a small but growing field of research [7, 9, 146, 153, 298, 341]. However, the contestable AI literature lacks guidance for application in specific design situations. In general, designers need *examples* and *instructions* to apply a framework effectively [161, 214]. We, therefore, seek to answer the following questions: RQ1: What are the characteristics of a contestable public AI system? RQ2: What are the challenges facing the implementation of contestability in public AI?

We ground our work in the use of camera cars: human-driven vehicles equipped with image sensors used for *vehicular urban sensing* (VUS). The primary motivation for these systems is increased efficiency (cost reduction), for example, for parking enforcement. Outside of the densest urban areas, costs of traditional means of parking enforcement quickly exceed collected fees [240]. Ethical concerns over using camera cars for these and other purposes reflect those around smart urbanism more broadly: data is captured without consent or notice, and its benefits favor those doing the capturing, leading to reductionist views and overly technocratic decision-making [184].

In this paper, we explore the shape contestable AI may take in the context of local government public services, and we describe the responses of civil servants who work with AI to these future visions.

Our design methods are drawn from speculative, critical, and future-oriented approaches [18, 90, 119, 191]. We use the 'Contestable AI by Design' framework [7] as a generative tool to design a concept for a contestable camera car system. Using the resulting concept video as a prompt, we conduct semi-structured interviews with civil servants who work with AI and are employed by the City of Amsterdam. Our focus here is on the challenges our respondents see towards implementing these future visions and contestability more generally. We then use reflexive thematic analysis [43–45] to generate themes from the interview transcripts that together describe the major challenges facing the implementation of contestability in public AI.[1]

The empirical work for this study was conducted in Amsterdam. The city has previously explored ways of making camera cars more "human-friendly."

1. This study was preregistered at Open Science Framework: `https://osf.io/26rts`

But efforts so far have been limited to up-front design adjustments to camera cars' physical form.[2]

The contributions of this paper are twofold: First, we create an example near-future concept of a contestable AI system in the context of public AI, specifically camera-based VUS. The concept video is usable for debating the merits of the contestable AI concept and exploring implications for its implementation. Second, we offer an account of the challenges of implementing contestability in public AI, as perceived by civil servants employed by Amsterdam who work with AI.

We structure this paper as follows: First, we introduce Amsterdam and its current use of camera cars for parking enforcement and other purposes. Next, we discuss related work on contestable AI, public and urban AI, VUS, and speculative design. Subsequently, we describe our research approach, including our design process, interview method, and data analysis. We then report on the resulting design concept and civil servant responses. Finally, we reflect on what our findings mean for current notions of contestable AI and consider the implications for its design in the context of public and urban AI in general and camera-based VUS in particular.

4.2 Background

4.2.1 Amsterdam

Amsterdam is the capital and largest city of the Netherlands. Its population is around 0.9 million (881.933 in 2022).[3] "By Dutch standards, the city is a financial and cultural powerhouse" [287].

Amsterdam is intensely urbanized. The city covers 219.492 km2 of land (2019). The city proper has 5.333 (2021) inhabitants per km2 and 2.707 (2019) houses per km2.[4] Amsterdam is considered the financial and business capital of the country. It is home to a significant number of banks and corporations. Its port is the fourth largest in terms of sea cargo in Northwest Europe.[5] Amsterdam is also one of the most popular tourist destinations in Europe.[6]

2. https://responsiblesensinglab.org/projects/scan-cars
3. https://onderzoek.amsterdam.nl/interactief/kerncijfers
4. https://onderzoek.amsterdam.nl/interactief/kerncijfers
5. https://www.amsterdam.nl/bestuur-organisatie/volg-beleid/economie/haven
6. https://onderzoek.amsterdam.nl/publicatie/bezoekersprognose-2022-2024

In 2022, over a third (35%) of residents were born abroad.[7] Amsterdam has relatively many households with a very low income (17%) and a very high income (14%).[8] In 2020, Amsterdam's working population (age 15–74) was relatively highly educated (48%).[9]

The city is governed by a directly elected municipal council, a municipal executive board, and a government-appointed mayor. The mayor is a member of the board but also has individual responsibilities. The 2022–2026 coalition agreement's final chapter on "cooperation and organization" contains a section on "the digital city and ICT," which frames technology as a way to improve services and increase equality and emancipation. Among other things, this section focuses on protecting citizens' privacy, safeguarding digital rights, monitoring systems using an algorithm register[10], testing systems for "integrity, discrimination and prejudice" throughout their lifecycle, and the continuing adherence to principles outlined in a local manifesto describing values for a responsible digital city[11].

4.2.2 Camera Car Use in Amsterdam

In January 2021, 13 municipalities in the Netherlands, including Amsterdam, made use of camera cars for parking monitoring and enforcement.[12]

Paid parking targets parking behavior and car use of citizens, businesses, and visitors. Its aims are to reduce the number of cars in the city, relieve public space pressures, and improve air quality. Cities expect to make alternative modes of transportation (cycling, public transport) more attractive by charging parking fees and limiting the availability of parking licenses per area.

The system in Amsterdam checks if parked cars have paid their parking fee or have a parking permit. Community service officers use cars outfitted with cameras to patrol city parking areas. They capture images of license plates and use computer vision algorithms to recognize license plates. The system uses these license plates to check with a national parking register if a vehicle has the right to park in its location and at the given time. Payment must be

7. https://onderzoek.amsterdam.nl/interactief/dashboard-kerncijfers
8. https://onderzoek.amsterdam.nl/publicatie/de-staat-van-de-stad-amsterdam-xi-2020-2021
9. https://onderzoek.amsterdam.nl/publicatie/de-staat-van-de-stad-amsterdam-xi-2020-2021
10. https://algoritmeregister.amsterdam.nl
11. https://tada.city
12. https://www.rtlnieuws.nl/nieuws/nederland/artikel/5207606/scanauto-boete-aanvechten-grote-steden-amsterdam-utrecht-den-haag

made within 5 minutes after the vehicle has been 'scanned.' If not, a parking inspector employed by the company that operates the system on behalf of the city reviews the situation based on four photos to determine if exceptional circumstances apply (e.g., curbside (un)loading, stationary at traffic light). This human reviewer also checks if the license plate is recognized correctly. In case of doubt, they dispatch a parking controller by motor scooter to assess the situation on-site. The system issues a parking fine if no exceptional circumstances apply by passing the data to the municipal tax authorities. They then use the same parking register database to retrieve the personal data of the owner of the vehicle to send them a parking fine.

A dedicated website allows people to appeal a fine within six weeks of issuing. The website provides access to the environment and license plate photos. (Any bystanders, unrelated license plates, and other privacy-sensitive information are made unrecognizable.) A third-party service also offers to object to traffic and parking fines on behalf of people, free of charge.

Amsterdam also uses parking monitoring camera cars to detect stolen vehicles and vehicles with a claim from the police or the public prosecutor. Cars are registered as stolen in the parking register. In case of a match with a scanned license plate, a national vehicle crime unit, possibly cooperating with the police, can take action. Data about 'parking pressure' and the types of license holders for municipal policy development is also collected.

Finally, Amsterdam is exploring additional applications of camera cars, including outdoor advertisement taxes[13] and side-placed garbage collection.[14]

4.3 Related Work

4.3.1 Contestable AI by Design

A small but growing body of research explores the concept of contestable AI [7, 9, 146, 153, 298, 341]. Contestability helps to protect against fallible, unaccountable, unlawful, and unfair automated decision-making. It does so by ensuring the possibility of human intervention throughout the system lifecycle and by creating arenas for adversarial debate between decision subjects and system operators.

13. https://responsiblesensinglab.org/projects/scan-cars
14. https://medium.com/maarten-sukel/garbage-object-detection-using-pytorch-and-yolov3-d6c4e0424a10

Hirsch et al. [153] define contestability as "humans challenging machine predictions," framing it as a way to protect against inevitably fallible machine models by allowing human controllers to intervene before machine decisions are put into force. Vaccaro et al. [341] frame contestability as a "deep system property," representing joint human-machine decision-making. Contestability is a form of procedural justice, giving voice to decision subjects and increasing perceptions of fairness. Almada [9] defines contestability as the possibility for "human intervention," which can occur not only post-hoc, in response to an individual decision, but also ex-ante, as part of AI system development processes. For this reason, they argue for a practice of "contestability by design." Sarra [298] argues that contestability exceeds mere human intervention. They argue that contestability requires a "procedural relationship." A "human in the loop" is insufficient if there is no possibility of a "dialectical exchange" between decision subjects and human controllers. Finally, Henin and Le Métayer [146] argue that the absence of contestability undermines systems' legitimacy. They distinguish between explanations and justifications. The former are descriptive and intrinsic to the systems themselves. The latter are normative and extrinsic, depending on outside references for assessing outcomes' desirability. Because contestability seeks to show that a decision is inappropriate or inadequate, it requires justifications in addition to explanations.

Building on these and other works, Alfrink et al. [7] define contestable AI as "open and responsive to human intervention, throughout their lifecycle, establishing a procedural relationship between decision subjects and human controllers." They develop a preliminary design framework that synthesizes elements contributing to contestability identified through a systematic literature review. The framework comprises five system features and six development practices, mapped to major system stakeholders and typical AI system lifecycle phases. For Alfrink et al. [7], contestability is about "leveraging conflict for continuous system improvement."

Most of the works Alfrink et al. [7] include are theoretical rather than empirical and are not derived from specific application contexts. Contexts that do feature in works discussed are healthcare [153, 278], smart cities [172], and content moderation [95, 342, 343]. The framework has not been validated and lacks guidance and examples for ready application by practitioners.

4.3.2 Public and Urban AI

An increasing number of researchers report on studies into the use of AI in the public sector, i.e., *public AI* [48, 86, 102, 106, 109, 230, 277, 300, 301, 346].

Although some do use the term "AI" [86, 102, 106], more commonly the term used is "algorithm" or "algorithmic system" [48, 109, 277, 300, 301, 346]. These algorithmic systems are put to use for informing or automating (public) decision-making by government public service (or sector) agencies [48, 86, 102, 300, 301]. The application contexts researchers report on include: child protection [48, 86, 300, 301, 346]; public housing [86]; public health [86, 277]; social protection [86, 109, 346]; public security [102, 230] and taxation [346]. Some of the issues explored include: how transparency, explanations and justifications may affect citizens' trust, acceptance and perceived legitimacy of public AI [48, 86, 106]; the politics of measurement, the human subjective choices that go into data collection, what does and does not get counted, and in what way [230, 277]; and how public sector employees' work is impacted by public AI [109], with a particular focus on discretion [300, 301], and how research and practice might more productively collaborate [346].

An overlapping but distinct area of research focuses on the role of AI in the built environment, so-called *urban AI* [6, 163, 219, 220, 299, 336, 363]. Many application contexts here are mobility-related, for example, smart electric vehicle charging [6]; autonomous vehicles [219]; and automated parking control systems [299]. The focus of this research tends to be more on how AI molds, mediates, and orchestrates the daily lived experience of urban places and spaces. Ethical questions related to AI become intertwined with city-making ethics, "who has the right to design and live in human environments" [220]. What the urban AI 'lens' adds to public AI discourse are questions of *spatial justice* [314] in addition to those of procedural and distributive justice.

4.3.3 Vehicular Urban Sensing

Vehicular (urban) sensing is when "vehicles on the road continuously gather, process, and share location-relevant sensor data" [205]. They are "a prominent example of cyber-physical systems" requiring a multidisciplinary approach to their design [273]. Sensors can be mounted on vehicles, or onboard smartphones may be used instead or in addition [101, 205]. Vehicles, here, are usually cars (automobiles). One advantage of cars is that they have few power constraints [273]. Much of the literature to date focuses on enlisting privately owned vehicles in crowdsourcing efforts [101, 201, 273, 373], as well as networking infrastructure challenges [49, 205, 273, 373]. A wide range of sensors is discussed, but some focus specifically on the use of cameras [32, 49, 240, 370]. Applications include traffic monitoring and urban surveillance [49], air pollution and urban traffic [273], infrastructure monitoring (i.e., "remote assessment of struc-

tural performance") [32], and (of particular note for our purposes here) parking monitoring and enforcement [240]. Mingardo [240] describes enforcement of on-street parking in Rotterdam, the Netherlands, using "scan-cars." They claim the main reason for introducing this system was to reduce the cost of enforcement. Income usually covers enforcement costs in areas with high fees and large numbers of motorists. However, residents usually have affordable parking permits in peripheral areas, and the area to cover is much larger. Systems like the one in Rotterdam use so-called "automatic number plate recognition" (ANPR). Zhang et al. [370] propose an approach to segmenting license plates that can deal with a wide range of angles, lighting conditions, and distances. They report an accuracy of 95%.

4.3.4 Speculative Design

We use 'speculative design' as a cover term for various forms of design futuring, including design fiction and critical design. Speculative design seeks to represent or "project" future consequences of a current issue [74].

Although early exemplars of speculative design often took the form of products, later projects usually include various forms of storytelling, primarily to aid audience interpretation and engagement [119]. Auger [16] calls this a design's "perceptual bridge." Sterling [319] frames design fiction as a marriage of science-fiction literature and industrial product design, which should address the inabilities of both to "imagine effectively." Kirby [183] has described the relationship between science-fiction cinema and design. Design in service of cinema produces "diegetic prototypes," objects that function within a film's story world. Alternatively, as Bleecker [31] puts it, speculative design produces things that tell stories and, in the audience's minds, create future worlds. This notion is similar to what Dunne and Raby [90] call "design as a catalyst for social dreaming." For them, the focus of speculation is on the implications of new developments in science and technology. As such, they claim speculative design can contribute to new "sociotechnical imaginaries" [170, 171].

Speculative design can be a way to "construct publics" around "matters of concern" [31, 74, 110], to "design for debate" [229]. It is about asking questions rather than solving problems [110, 119]. To spark debate, speculative design must be provocative [19]. It evokes critical reflection using satirical wit [227]. For this satire to work, the audience must read speculative designs *as* objects of design, contextualized and rationalized with a narrative of use [227, 229]. Speculative designs do not lack function and can, therefore, not be dismissed as mere art. Instead, speculative design leverages a broader conception of function

that goes beyond traditional notions of utility, efficiency, and optimization and instead seeks to be relational and dynamic [229].

To further support audiences' engagement in debate, some attempts have combined speculative design with participatory approaches. In workshop-like settings, speculative designs co-created with audiences can surface controversies and be a form of "infrastructuring" that creates "agonistic spaces" [110, 119, 151].

Early work was primarily focused on speculative design as a 'genre,' exploring what designs can do, and less on how it should be practiced [119]. Since then, some have explored speculative design as a method in HCI design research, particularly in 'research through design' or 'constructive design research' [17, 19, 119].

There have been a few attempts at articulating criteria by which to evaluate speculative designs [17, 18, 74, 119]. Some works offer guidelines for what makes speculative design critical [17]; reflecting on speculative designs [195]; evaluations that match expected knowledge outcomes [21]; and 'tactics' for that drawn from a canon of exemplars [104].

4.4 Method

Our overall approach can be characterized as constructive design research that sits somewhere between what Koskinen et al. [194] calls the 'field' and 'showroom' modes or research through design using the 'genre' of speculative design [119]. We create a concept video of a near future contestable camera car. We actively approach our audience to engage with the concept video through interviews. We use storytelling to aid audience interpretation, to help them recognize how a contestable camera car might fit into daily life. We seek to strike a balance between strangeness and normality. We measure success by the degree to which our audience is willing and able to thoughtfully engage with the concept video. In other words, we use speculative design to ask questions rather than provide answers.

Our study is structured as follows: (1) we first formulate a design brief to capture the criteria that the speculative design concept video must adhere to; (2) we then conduct the speculative design project; (3) a rough cut of the resulting concept video is assessed with experts; (4) the video is then adjusted and finalized; (5) using the final cut of the speculative design concept video as a 'prompt' we then conduct semi-structured interviews with civil servants; (6) finally, we use the interview transcripts for reflexive thematic analysis, exploring civil servants' views of challenges facing the implementation of contestability.

The data we generate consists of (1) visual documentation of the design concepts we create and (2) transcripts of semi-structured interviews with respondents. The visual documentation is created by the principal researcher and design collaborators as the product of the design stage. The transcripts are generated by an external transcriber on the basis of audio recordings.

4.4.1 Design Process

We first created a design brief detailing assessment criteria for the design outcomes, derived partly from Bardzell et al. [18]. The brief also specified an application context for the speculated near-future camera car: trash detection. We drew inspiration from an existing pilot project in Amsterdam. Garbage disposal may be a banal issue, but it is also multifaceted and has real stakes. We hired a filmmaker to collaborate with on video production. Funding for this part of the project came from AMS Institute, a public-private urban innovation center.[15] We first created a mood board to explore directions for the visual style. Ultimately, we opted for a collage-based approach because it is a flexible style that would allow us to depict complex actions without a lot of production overhead. It also struck a nice balance between accessibility and things feeling slightly off. We then wrote a script for the video. Here, we used contestability literature in general and the 'Contestable AI by Design' framework [7], in particular, to determine what elements to include. We tried to include a variety of risks and related system improvements (rather than merely one of each) so that the audience would not quickly dismiss things for lack of verisimilitude. Having settled on a script, we then sketched out a storyboard. Our main challenge here was to balance the essential depiction of an intelligent system with potential risks, ways citizens would be able to contest, and the resulting system improvements. As we collaboratively refined the storyboard, our filmmaker developed style sketches that covered the most essential building blocks of the video.[16] Once we were satisfied with the storyboard and style sketches, we transitioned into video production. Production was structured around reviews of weekly renders. On one occasion, this review included partners from AMS Institute. Our next milestone was to get a rough cut of the video 'feature complete' for assessment with experts.

For this assessment, we created an interview guide and a grading rubric. We based the rubric on the assessment criteria developed in the original design

15. https://www.ams-institute.org
16. Design brief, script, and storyboards are available as supplementary material.

brief. All experts were colleagues at our university, selected for active involvement in the fields of design, AI, and ethics. We talked to seven experts (five male, two female; two early-career researchers, three mid-career, and two senior). Interviews took place in early February 2022. Each expert was invited for a one-on-one video call of 30–45 minutes. After a brief introduction, we went over the rubric together. We then showed the concept video rough cut. Following this, the expert would give us the grades for the video. After this, we had an open-ended discussion to discuss potential further improvements. Audio of the conversations was recorded with informed consent and (roughly) transcribed using an automated service. We then informally analyzed the transcripts to identify the main points of improvement. We first summarized the comments of each respondent point by point. We then created an overall summary, identifying seven points of improvement. We visualized the rubric score Likert scale data as a diverging stacked bar chart.[17]

Once we completed the expert assessment, we identified improvements using informal analysis of the automated interview transcripts. The first author then updated the storyboard to reflect the necessary changes. We discussed these with the filmmaker and agreed on what changes were necessary and feasible. The changes were then incorporated into a final cut, adding music and sound effects created by a sound studio and a credits screen.

4.4.2 Civil Servant Interviews

Interviews were conducted from early May through late September 2022. We used purposive and snowball sampling. We were specifically interested in acquiring the viewpoint of civil servants involved in using AI in public administration. We started with a hand-picked set of five respondents, whom we then asked for further people to interview. We prioritized additional respondents for their potential to provide diverse and contrasting viewpoints. We stopped collecting data when additional interviews failed to generate significantly new information. We spoke to 17 respondents in total. Details about their background are summarized in Table 4.1. We invited respondents with a stock email. Upon expressing their willingness to participate, we provided respondents with an information sheet and consent form and set a date and time. All interviews were conducted online, using videoconferencing software. Duration was typically 30–45 minutes. Each interview started with an off-the-record introduction,

17. Interview guide, assessment form template, completed forms, tabulated assessment scores, and informal analysis report are available as supplementary material.

after which we started audio recording with informed consent from respondents. We used an interview guide to help structure the conversation but were flexible about follow-up questions and the needs of respondents. After a few preliminary questions, we would show the video. After the video, we continued with several more questions and always ended with an opportunity for the respondents to ask questions or make additions for the record. We then ended the audio recording and asked for suggested further people to approach. After each interview, we immediately archived audio recordings and updated our records regarding whom we spoke to and when. We then sent the audio recordings to a transcription service, which would return a document for our review. We would review the transcript, make corrections based on a review of the audio recording where necessary, and remove all identifying data. The resulting corrected and pseudonymized transcript formed the basis for our analysis.[18]

Table 4.1
Summary of civil servant interview respondent demographics.

Item	Category	Number
Gender	Female	10
	Male	7
Department	Digital Strategy and Information	3
	Legal Affairs	2
	Traffic, Public Space, and Parking	2
	Urban Innovation and R&D	10
Background	AI, arts & culture, business, data science, information science, law, philosophy, political science, sociology	–

4.4.3 Analysis

Our analysis of the data is shaped by critical realist [115, 136] and contextualist [147, 169] commitments. We used reflexive thematic analysis [43–45] because it is a highly flexible method that readily adapts to a range of questions, data generation methods, and sample sizes. Because of the accessibility of its results, it is also well-suited to our participatory approach. The principal researcher took the lead in data analysis. Associate researchers contributed with partial coding and review of coding results. The procedure for turning "raw" data into analyzable form was: (1) reading and familiarization; (2) selective coding (developing a corpus of items of interest) across the entire dataset; (3)

18. Interview guide is available as supplementary material.

searching for themes; (4) reviewing and mapping themes; and (5) defining and naming themes. We conducted coding using Atlas.ti. We used a number of credibility strategies: member checking helped ensure our analysis reflected the views of our respondents; different researchers analyzed the data, reducing the likelihood of a single researcher's positionality overly skewing the analysis; and reflexivity ensured that analysis attended to the viewpoints of the researchers as they relate to the phenomenon at hand.[19] In what follows, all direct quotes from respondents were translated from Dutch into English by the first author.

4.5 Results

4.5.1 Concept Video Description

The concept video has a duration of 1 minute and 57 seconds. Several stills from the video can be seen in Figure 4.1. It consists of four parts. The first part shows a camera car identifying garbage in the streets and sending the data off to an unseen place of processing. We then see the system building a heat map from identified garbage and a resulting prioritization of collection services. Then, we see garbage trucks driving off and a sanitation worker tossing the trash in a truck. The second part introduces three risks conceivably associated with the suggested system. The first risk is the so-called 'chilling effect.' People feel spied on in public spaces and make less use of it. The second risk is the occurrence of 'false positives,' when objects that are not garbage are identified as such, leading to wasteful or harmful confrontations with collection services. The third risk is 'model drift.' Prediction models trained on historical data become out of step with reality on the ground. In this case, collection services are not dispatched to where they should be, leading to inexplicable piling up of garbage. The third part shows how citizens introduced in the risks section contest the system using a four-part loop. First, they use explanations to understand system behavior. Second, they use integrated channels for contacting the city about their concern. Third, they discuss their concern and points of view with a city representative. Fourth, the parties decide on how to act on the concern. The fourth and final part shows how the system is improved based on contestation decisions. The chilling effect is addressed by explicitly calling out the camera car's purpose on the vehicle itself, and personal data is discarded before transmission. False positives are guarded against by having a human controller review images that the system predicts contain trash before action is taken. Finally, model drift is

19. Interview transcript summaries and codebook are available as supplementary material.

prevented by regularly updating models with new data. The video ends with a repeat of garbage trucks driving off and a sanitation worker collecting trash. A credits screen follows it.[20]

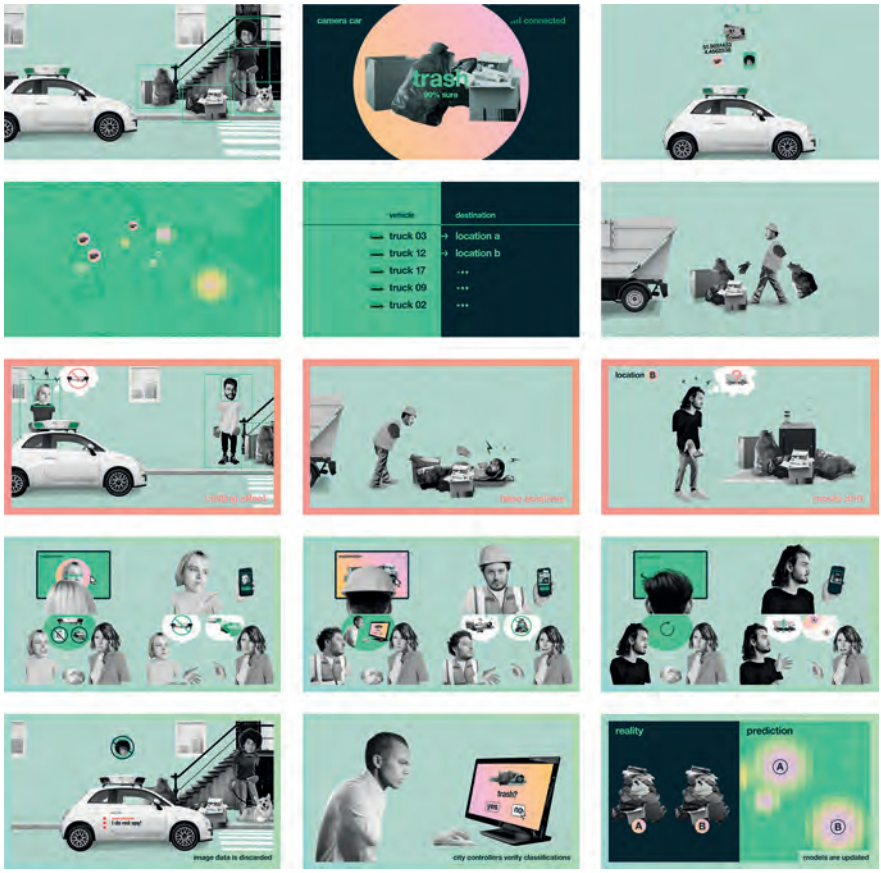


Figure 4.1
Stills from concept video.

20. The concept video is available as supplementary material.

4.5.2 Civil Servant Responses to Concept Video

From our analysis of civil servant responses to the concept video, we constructed three themes covering 13 challenges. See Table 4.2 for a summary.

Table 4.2
Overview of themes and associated challenges.

Theme	#	Challenge
Enabling civic participation (4.5.2)	T1.1	Citizen capacities
	T1.2	Communication channels
	T1.3	Feedback to development
	T1.4	Reporting inequality
	T1.5	Participation limitations
Ensuring democratic embedding (4.5.2)	T2.1	Democratic control
	T2.2	External oversight
	T2.3	Dispute resolution
Building capacity for responsibility (4.5.2)	T3.1	Organizational limits
	T3.2	Accountability infrastructure
	T3.3	Civil servant capacities
	T3.4	Commissioning structures
	T3.5	Resource constraints

Theme 1: Enabling Civic Participation

T1.1 *Citizen capacities* (P1, P4, P5, P9, P10, P11, P12, P16, P17): Several respondents pointed out that contestability assumes sovereign, independent, autonomous, empowered, and articulate citizens. Citizens need sufficient awareness, knowledge, and understanding of systems to contest effectively.

> But everything actually starts with that information position as far as I am concerned. (P10)

It can be hard for people to understand the metrics used for evaluating model performance. For example, P17 describes how a model's intersection over union (IOU) score of 0.8 was talked about internally as an accuracy score of 80%. Individuals also struggle to identify systemic shortcomings. Their view is limited to the impacts directly relating to themselves only. They may not even be aware that the decision that has impacted them personally was made in part by an algorithm. In addition, citizens can have false views of what systems do. For example, citizens and civic groups believed parking enforcement camera cars recorded visual likenesses

of people in the streets, which was not the case. Citizens' ability to effectively contest further depends on how well they can navigate the city government's complicated internal organizational structure.

Many respondents describe how citizens' willingness to engage depends on their view of city government. Those who feel the city does not solve their problems will be reluctant to participate. Citizens' inclination to scrutinize public algorithmic systems also depends on their general suspicion of technology. This suspicion appears to be at least somewhat generational. For example, younger people are more cautious about sharing their data. Suspicion is contextual, depending on what is at stake in a given situation. A lack of trust can also lead to citizens rejecting explanations and justifications offered by the city.

> I just think what a challenge it is to have a substantive conversation and how do you arrive at that substantive conversation. (P16)

T1.2 *Communication channels* (P3, P4, P7, P8, P11, P12, P14, P16): Many respondents recognize the importance of ensuring citizens can talk to a human representative of the city. Currently, citizens can contact the city about anything using a central phone number. Reports from citizens are subsequently routed internally to the proper channels.

Ideally, the city should be able to route questions related to AI to civil servants who understand the relevant systems. Citizens are not able nor responsible for determining which issues pertain to algorithms and which do not. Triage should happen behind the scenes, as is currently the case with the central phone line. In other words, respondents would not favor a separate point-of-contact for 'digital matters.'

Executive departments are responsible for work processes, including those that use AI. They should, therefore, be the ones answering questions, *including* those that relate to technology. But this is currently not always the case. Some respondents point out that development teams cannot be made responsible for answering citizens' questions. Despite this fact, these respondents describe how their development teams *do* receive emails from citizens and simply answer them.

Beyond a central phone line, some respondents are considering other easily accessible, lower-threshold interaction modalities for expressing disagreement or concern (cf. Item T2.3).

T1.3 *Feedback to development* (P1, P2, P3, P4, P5, P7, P10, P13, P14, P15, P17): Respondents feel it is important for development teams to seek feedback from citizens during development. Indeed, for those systems developed internally, it is currently common practice to follow some iterative develop-

ment methodology that includes testing pre-release software with citizen representatives. Most of the algorithmic systems discussed by respondents are still in this so-called pilot stage. Pilots are used to test new ideas for viability and explore the practical and ethical issues that might arise when a system is taken into regular everyday operation.

> But I also think testing is necessary for these kinds of things. So, if you think it through completely, you will eventually see if you test whether it is feasible. Because now I have every time with such an iteration [...] you run into other things that make you think, how is this possible? (P12)

The city also conducts pilots to identify what is needed to justify the use of technology for a particular purpose.

> So we start a pilot in the situation where we already think: we have to take many measures to justify that. Because bottom line, we think it is responsible, but what do you think about this if we do it exactly this way? Do you agree, or is that [...] do you use different standards? (P7)

Respondents involved with system development recognize that feedback from citizens can help eliminate blind spots and may lead to new requirements.

Some respondents argue that all reports received by algorithmic system feedback channels should be open and public, or at least accessible to the municipal council so that democratic oversight is further enabled (cf. Item T2.1).

On a practical level, to close the loop between citizens' reports and development, infrastructure is needed (cf. Item T3.2). For example, the city's service management system, which integrates with the internal software development environment, is not yet open to direct reports from citizens but only from human controllers (cf. Item T3.2). For those systems using machine learning models, there are no provisions yet for capturing feedback from citizens to retrain models (e.g., in a supervised learning approach).

T1.4 *Reporting inequality* (P1, P4, P6, P12, P14, P15): Several respondents mentioned the issue of "reporting inequality," where some citizens are more able and inclined to report issues to the city than others (cf. Item T1.1). Some recent VUS efforts aim to counteract this reporting inequality; for example, the trash detection pilot our concept video took as a source of inspiration. Affluent neighborhoods are known to report on stray trash more than disadvantaged areas do and, as a result, are served better than is considered fair.

Because of reporting inequality, respondents are weary of approaches that tie system changes directly to individual reports. For example, contestability may counteract the unequal distribution of vehicles due to system flaws, but it may just as well reintroduce the problem of reporting inequality. Contestability runs the risk of giving resourceful citizens even more outsize influence. Other respondents counter that making system changes in response to individual complaints may still be warranted if those changes benefit most citizens.

Ultimately, many respondents feel it is up to developers and civil servants to interpret and weigh the signals they receive from citizens (cf. Item T1.3).

T1.5 *Participation limitations* (P1, P2, P4, P5, P6, P8, P10, P12, P14, P15, P16): Just as governments should be aware of reporting inequality (cf. Item T1.4), they should also ensure participation and contestation are representative. A real risk is that those with technical know-how and legal clout shape the debate around algorithmic systems. Respondents repeatedly point out that existing citizen participation efforts struggle to ensure diversity, inclusion, and representation.

> For example, in [district], we also met someone who did many development projects with the neighborhood and who also agreed that, of course, the empowered people or the usual suspects often provide input, and in [district], also low literacy and all sorts of other things make it much more difficult to [...] provide input if it is their neighborhood [...]. (P2)

For the city, it is a struggle to find citizens willing and able to contribute to participation processes. Sometimes, as a solution, the city compensates citizens for participating. Another way to improve inclusion is to go where citizens are rather than expect them to approach the city—for example, by staging events and exhibitions as part of local cultural festivals or community centers.

Participation efforts assume direct representation. There is no mechanism by which individuals can represent interest groups. Citizens do not represent anyone but themselves and are not legally accountable for their decisions. Respondents point out that as one goes up the participation ladder [15, 56] more obligations should accompany more influence.

Some respondents point out that governments should take responsibility and depend less on individual citizens or hide behind participatory processes.

Theme 2: Ensuring Democratic Embedding

T2.1 *Democratic control* (P1, P2, P3, P4, P5, P6, P7, P9, P10, P11, P12, P13, P14, P15, P16, P17): Several respondents pointed out that the discretion to use AI for decision-making lies with the executive branch. For this reason, the very decision to do so, and the details of *how* an algorithmic system will enact policy, should, in respondents' eyes, be a political one. Debate in the municipal council about such decisions would improve accountability.

Respondents identify a tension inherent in public AI projects: Policy-makers (alderpersons) are accountable to citizens and commission public AI projects, but they often lack the knowledge to debate matters with public representatives adequately. On the other hand, those who build the systems lack accountability to citizens. Accountability is even more lacking when developers do not sit within the municipal organization but are part of a company or non-profit from which the city commissions a system (cf. Item T3.4).

Respondents also point out that contestations originate with individual citizens or groups but also with elected representatives. In other words, the municipal council *does* monitor digital developments. The legislature can, for example, shape how the executive develops AI systems by introducing policy frameworks.

P7 outlined three levels of legislation that embed municipal AI projects: (a) the national level, where the city must determine if there is indeed a legal basis for the project; (b) the level of local ordinances, which ideally are updated with the introduction of each new AI system so that public accountability and transparency are ensured; and finally (c) the project or application level, which focuses on the 'how' of an AI system, and in the eyes of P7 is also the level where direct citizen participation makes sense and adds value (cf. Item T1.5).

Feedback on AI systems may be about business rules and policy, which would require a revision *before* a technical system can be adjusted.[21] This then may lead to the executive adjusting the course on system development under its purview (cf. Item T1.3).

There is also an absence of routine procedures for reviewing and updating existing AI systems in light of the new policy. Political preferences of elected city councils are encoded in business rules, which are translated into code. Once a new government is installed, policy gets updated, but

21. This entanglement of software and policy is well-described by Jackson et al. [168].

related business rules and software are not, as a matter of course, but should be.

T2.2 *External oversight* (P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P17): The city makes use of several forms of external review and oversight. Such reviews can be a requirement or something the city seeks out because of, for example, citizens' lack of trust (cf. Item T1.1).

A frequently mentioned body is the local Personal Data Commission (PDC). A PDC review is mandatory when a prospective algorithmic system processes personal data or when it is considered a high-risk application. The PDC focuses, among other things, on a system's legal basis, proportionality, and mitigation of identified risks. One respondent proposes that such human rights impact assessments be made open for debate.[22]

Other review and oversight bodies include the local and national audit offices, the municipal ombudsperson, and a so-called reporting point for chain errors. One shortcoming is that many of these are incident-driven. They cannot proactively investigate systems.

Naturally, the civil servants, committee members, ombudspersons, and judges handling such cases must have a sufficient understanding of the technologies involved. External review bodies sometimes, at least in respondents' eyes, lack sufficient expertise. One example of such a case is recent negative advice delivered by a work participation council after a consultation on using AI by the work participation and income department to evaluate assistance benefit applications. At least one respondent involved in developing the system proposal felt that, despite considerable effort to explain the system design, the council did not fully grasp it.

22. Following widespread resistance against a 1971 national census, the Dutch government established a commission in 1976 to draft the first national privacy regulation. Because it collected and processed a significant amount of personal data itself, Amsterdam decided not to wait and created local regulations in 1980. Every municipal service and department was required to establish privacy regulations. The city established a special commission to review these guidelines and to decide if municipal bodies were allowed to exchange information, thereby creating the PDC ("Commissie Persoonsgegevens Amsterdam (CPA)," `https://assets.amsterdam.nl/publish/pages/902156/brochure_cpa_40_jarig_bestaan.pdf`). The executive board expanded the tasks of the PDC in December 2021 (`https://www.amsterdam.nl/bestuur-organisatie/college/nieuws/nieuws-19-januari-2022/`). It now advises the board, upon request or on its own initiative, on issues "regarding the processing of personal data, algorithms, data ethics, digital human rights and disclosure of personal data" (`https://assets.amsterdam.nl/publish/pages/902156/cpa_reglement.pdf`). In the lead-up to this decision, in April 2021, a coalition of green, left, and social liberal parties submitted an initiative proposal to the board that aimed to "make the digital city more humane." It, too, argued for the expansion of the PDC's role (`https://amsterdam.groenlinks.nl/nieuws/grip-op-technologie`).

P7 considers judicial review by an administrative court of a decision that is at least in part informed by an algorithmic system, the "finishing touch." When a client file includes data that significantly impacts a model prediction, a judge's ruling on a municipal decision is implicitly also about the operation of the model.

> If [a decision] affects citizens in their legal position, for example, in the case of a fine [...], then yes, the administrative court can look into it. That is when it gets exciting. That is the finishing touch to what we have come up with. (P7)

This sentiment was echoed by P11 when they discussed how they could show in court what images the municipal parking monitoring camera car exactly captured, which received a favorable ruling from a judge.

T2.3 *Dispute resolution* (P10, P11, P14, P15): Respondents feel that, for individual substantive grievances caused by algorithmic decision-making, existing complaint, objection, and appeal procedures should also work. These form an escalating ladder of procedures: complaints are evaluated by civil servants; objections go to an internal committee; if these fail, the case is handled by an ombudsperson; and finally, appeals procedures are handled by a judge.

Respondents point out that existing procedures can be costly and limiting for citizens and not at all "user-friendly." Existing procedures still rely heavily on communication by paper mail. Current procedures can be stressful because people are made to feel like an offender rather than being given the benefit of the doubt.

> And we criminalize the citizen very quickly if he does not want to—a difficult citizen, annoying. Yes, no, it is just that way, and no, sorry, bye. So there is little to no space, and if you have heard [a complaint] ten times from citizens, then maybe you should think about, we have ten complaining citizens. It is not one or two. There might be something wrong, so let us look at that. (P13)

Respondents agree that more effort should be put into creating alternative dispute resolution mechanisms. These should help citizens stay out of costly and stressful legal proceedings. However, these ideas are mostly considered an 'innovation topic,' which is to say, it is not part of daily operation. Such measures would require collaboration between those departments executing work processes and legal. At the moment, execution tends to consider dealing with disputes as not part of their remit. Legal does currently call citizens who have started an appeals procedure to make

them feel heard, find alternative solutions, and offer them the opportunity to withdraw.

Existing mechanisms do require more integration with technology. For example, case files should include all the relevant information about the data and algorithms used. Some services, such as parking monitoring, have already built custom web interfaces for appeals that integrate with algorithmic systems and offer citizens access to their case data. These would either expedite otherwise unwieldy legacy procedures or seek to keep citizens out of formal legal appeal procedures altogether.

Theme 3: Building Capacity for Responsibility

T3.1 *Organizational limits* (P4, P5, P7, P11, P15): Respondents point out that organizational fragmentation works against the city's capacity to respond to citizen reports. The problem is not necessarily that signals are not received by the city. Often, the problem is that they are not adequately acted upon. Internal fragmentation also makes it hard for citizens to know whom they should approach with questions (cf. Items T1.1, T1.2). For example, with parking, citizens are inclined to go to their district department, and these need to pass on questions to parking enforcement, who, in turn, if it concerns a street-level issue, must dispatch a community service officer.

> And I think that if you cut up the organization as it is now [...], then you might also have to work with other information in order to be able to deliver your service properly. So when we all had [more self-sufficient, autonomous] district councils in the past and were somewhat smaller, you could, of course, immediately say that this now has priority, we receive so many complaints, or the alderman is working on it. (P11)

Fragmentation and the bureaucratic nature of the city organization work against the adoption of 'agile methods.' Although pilots are in many ways the thing that makes the innovation funnel of the city function, respondents also describe pilots as "the easy part." The actual implementation in daily operations is a completely different matter. P3 describes this as the "innovation gap." Transitioning a successful pilot into operation can easily take 3–5 years.

T3.2 *Accountability infrastructure* (P2, P4, P5, P7, P11, P12, P13): Respondents discuss various systems that are put in place to improve accountability. The city is working to ensure requirements are traceable back to the person who set them, and developers record evidence to show they are met.

Evidence would include email chains that record design decisions and system logging that shows specific measures are indeed enforced (such as deletion of data). Regarding models, respondents indicate the importance of validating them to demonstrate that they indeed do what they are said to do.

Once past the pilot stage, monitoring and maintenance become essential considerations currently under-served. For this purpose, developers should correctly document systems in anticipation of handover to a maintenance organization. Systems must be ensured to operate within defined boundaries, both technical and ethical (impact on citizens), and the delivery of "end-user value" must also be demonstrated. Such monitoring and maintenance in practice require the system developers' continued involvement for some time.

Another provision for accountability is the service management system integrated with the city's software development and operations environment (cf. Item T1.3). Several respondents pointed out that surveillance and enforcement are two separate organizational functions. For those AI systems related to surveillance and enforcement, a 'human-in-the-loop' is currently already a legal requirement at the enforcement stage. Human controllers use the service management system to report system flaws, which may lead to changes and are fully traceable (cf. Item T1.3). Once in maintenance, with these systems in place, it should be possible for functional management to revise systems periodically, also in light of policy changes (cf. Item T2.1).

Several respondents argue that the city should also monitor individual complaints for issues that require a system change (cf. Item T1.3).

T3.3 *Civil servant capacities* (P1, P3, P4, P6, P7, P15): Contestability puts demands on civil servants.

> [...] I think all contestability [shown in the video] assumes a very assertive citizen who is willing to contact a city representative who is willing to listen, has time for it, and is committed to doing something about it. (P1)

Civil servants need knowledge and understanding of AI systems, including those employees who speak to citizens who contact the city with questions, e.g., through the central phone number. Politicians, city council members, and alderpersons also need this understanding to debate the implications of new systems adequately. At the level of policy execution, department heads and project leads are the "first line of defense" when things go wrong (P7). So, they cannot rely on the expertise of development teams but must have a sufficient understanding of matters themselves.

Finally, legal department staff must also understand algorithms. P15 mentions that a guideline is being made that should aid in this matter.

Beyond updating the knowledge and skills of existing roles, new roles are necessary. In some cases, agile-methods-style 'product owners' act as those who translate policy into technology. However, P7 feels the organization as a whole still lacks people who can translate legislation and regulations into system requirements. Zooming out further, respondents mention challenges with the current organizational structure and how responsibility and accountability require multidisciplinary teams that can work across technical and social issues (cf. Item T3.1).

T3.4 *Commissioning structures* (P1, P3, P4, P11, P12, P13, P16, P17): The city can commission AI systems in roughly three ways, with different impacts on the level of control it has over design, development, and operation: (a) by purchasing from a commercial supplier a service that may include an AI system; (b) by outsourcing policy execution to a third party, usually a non-profit entity who receives a subsidy from the city in return; or (c) by developing a system in-house.

When purchasing, the city can exercise control mainly by imposing purchasing conditions, requiring a strong role as a commissioner. When out-placing policy execution, the city has less control but can impose conditions on the use of technology as part of a subsidy provision. When developing in-house, the city owns the system completely and is therefore in full control. In all cases, however, the city is the 'policy owner' and remains responsible for executing the law.

These different collaboration structures also shape the possible dialogue between policy-makers and system developers at the start of a new project. When development happens in-house, an open conversation can happen. In the case of a tender, one party cannot be advantaged over others, so there is little room for hashing things out until an order is granted.

Of course, collaboration with external developers can also have "degrees of closeness" (P4). More or less 'agile' ways of working can be negotiated as part of a contract, which should allow for responding to new insights mid-course.

Purchasing managers sometimes perceive what they are doing as the acquisition of a service that is distinct from buying technology solutions and can sometimes neglect to impose sufficient conditions on a service provider's use of technology.

The duration of tenders is typically three years. On occasion, the city comes to new insights related to the responsible use of technologies a

service provider employs (e.g., additional transparency requirements). However, it cannot make changes until after a new tender. Respondents point out that an additional feedback loop should lead to the revision of purchasing conditions. P17 describes a project in which parts of the development and operation are outsourced, and other components are done in-house. The decision on what to outsource mainly hinges on how often the city expects legislature changes that demand system updates.

T3.5 *Resource constraints* (P3, P4, P12, P16, P17): Supporting contestability will require additional resource allocation. Respondents point out that the various linchpins of contestable systems suffer from limited time and money: (a) conducting sufficiently representative and meaningful participation procedures; (b) having knowledgeable personnel available to talk to citizens who have questions or complaints; (c) ensuring project leads have the time to enter information into an algorithm register; (d) performing the necessary additional development work to ensure systems' compliance with security and privacy requirements; and (e) ensuring proper evaluations are conducted on pilot projects.

P12 compares the issue to the situation with freedom of information requests, where civil servants who are assigned to handle these are two years behind. Similarly, new legislation, such as the European AI Act, is likely to create even more work for the city.

For new projects, the city will also have to predict the volume of citizen requests so that adequate staffing can be put into place in advance. Having a face-to-face dialogue in all instances will, in many cases, be too labor-intensive (cf. Item T1.2). A challenge with reports from citizens is how to prioritize them for action by city services, given limited time and resources (cf. Item T1.4).

4.6 Discussion

Our aim has been two-fold: (1) to explore characteristics of contestable public AI and (2) to identify challenges facing the implementation of contestability in public AI. To this end, we created a speculative concept video of a contestable camera car and discussed it with civil servants employed by Amsterdam who work with AI.

4.6.1 Summary of Results

Concept Video: Example of Contestable Public AI

The speculative design concept argues for contestability from a risk mitigation and quality assurance perspective. First, it shows several hazards related to camera car use: chilling effect, false positives, and model drift. Then, it shows how citizens use contestability mechanisms to petition the city for system changes. These mechanisms are explanations, channels for appeal, an arena for adversarial debate, and an obligation to decide on a response. Finally, the video shows how the city improves the system in response to citizen contestations. The improvements include data minimization measures, human review, and a feedback loop back to model training. The example application of a camera car, the identified risks, and resulting improvements are all used as provocative examples, not as a prescribed solution. Together they show how, as Alfrink et al. [7] propose, "contestability leverages conflict for continuous system improvement."

Civil Servant Interviews: Contestability Implementation Challenges

From civil servant responses to the concept video, we constructed three themes:

- T1 *Enabling civic participation* (4.5.2): Citizens need skills and knowledge to contest public AI on equal footing. Channels must be established for citizens to engage city representatives in a dialogue about public AI system outcomes. The feedback loop from citizens back to system development teams must be closed. The city must mitigate against 'reporting inequality' and the limitations of direct citizen participation in AI system development.
- T2 *Ensuring democratic embedding* (4.5.2): Public AI systems are embedded in various levels of laws and regulations. An adequate response to contestation may require policy change before technology alterations. Oversight by city council members must be expanded to include scrutiny of AI use by the executive. Alternative non-legal dispute resolution approaches that integrate tightly with technical systems should be developed to complement existing complaint, objection, and appeal procedures.
- T3 *Building capacity for responsibility* (4.5.2): City organizations' fragmented and bureaucratic nature fights against adequately responding to citizen signals. More mechanisms for accountability are needed, including logging system actions and monitoring model performance. Civil servants

need more knowledge and understanding of AI to engage with citizens adequately. New roles that translate policy into technology must be created, and more multidisciplinary teams are needed. Contracts and agreements with external development parties must include responsible AI requirements and provisions for adjusting course mid-project. Contestability requires time and money investments across its various enabling components.

Diagram: Five Contestability Loops

We can assemble five contestability loops from civil servants' accounts (Figure 4.2). This model's backbone is the primary loop where citizens elect a city council and (indirectly) its executive board (grouped as "policy-makers"). Systems developers translate the resulting policy into algorithms, data, and models. (Other policy is translated into guidance to be executed by humans directly.) The resulting "software," along with street-level bureaucrats and policy, form the public AI systems whose decisions impact citizens.

Our model highlights two aspects that are particular to the public sector context: (1) the indirect, representative forms of citizen control at the heart of the primary policy-software-decisions loop and (2) the second-order loops that monitor for systemic flaws that require addressing in upstream systems development or policy-making.

These five loops highlight specific intervention points in public AI systems. They indirectly indicate what forms of contestation could exist and between whom. To be fully contestable, we suggest that public AI systems implement all five loops. Better integration with the primary loop and the implementation of second-order monitoring loops deserve particular attention.

4.6.2 Results' Relation to Existing Literature

Contestable AI by Design

Following Alfrink et al. [7]'s definition of contestable systems as "open and responsive to human intervention," our respondents appear broadly sympathetic to this vision, particularly the idea that government should make more of an effort to be open and responsive to citizens.

We recognize many key contestability concepts in current city efforts as described by our respondents. For example, the possibility of human intervention [153] is mandatory in cases of enforcement, which can protect against

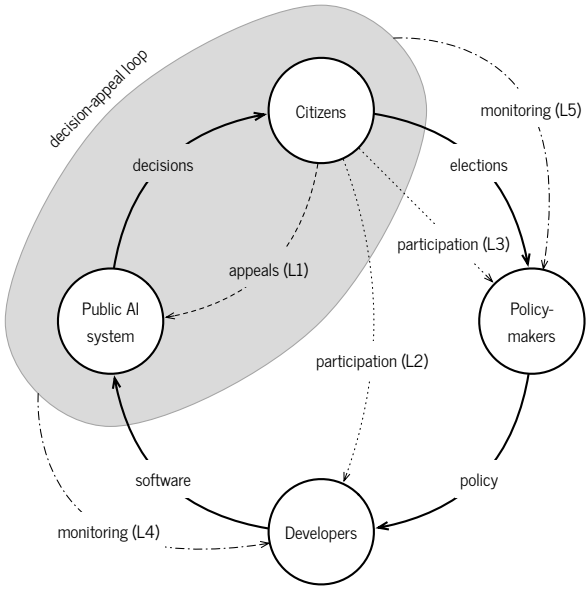


Figure 4.2
Diagram of our Five Loops Model, showing the basic flow of policy through software into decisions (solid arrows), the direct way citizens can contest individual decisions (L1, dashed arrow), the direct ways in which citizens can contest systems development and policy making (L2–3, dotted arrows), and the second-order feedback loops leading from all decision-appeal interactions in the aggregate back to software development and policy-making (L4–5, dashed-dotted arrows).

model fallibility, at least to the extent errors can be detected by individual human controllers. Nevertheless, this human-in-the-loop is implemented more for legal compliance than quality control. Respondents talk about quality assurance and ways to achieve it, e.g., through audits and monitoring, but few practical examples appear to exist as of yet. The city recognizes the need to integrate institutional contestability provisions with technical systems (i.e., contestability as "deep system property" [341]). However, this integration is currently underdeveloped. Positive examples include the custom web interface for appealing parking enforcement decisions. Ex-ante contestability measures [9] are present mainly in pilots in the form of civic participation in early-stage systems design. However, most participation happens on the project level and has no impact on policy decisions upstream from technology design. A dialectical relationship [298] is present on the far ends of what we could describe as the

question-complaint-object-appeal spectrum; for example, the central phone line on one end and the review of algorithmic decisions by administrative courts on the other. The middle range seems to have less opportunity for exchanging arguments; again, these measures generally lack integration with technology. In any case, executing this ideal at scale will be costly. Finally, the city appears to approach accountability and legitimacy by ensuring the availability of explanations (e.g., in the form of an algorithm register). There appears to be less interest in or awareness of, the need for justifications [146] of decisions.

Most of the literature emphasizes contestability from below and outside but does not account for the representative democracy mechanisms in which public AI systems are embedded. In terms of our Five Loops Model, city efforts emphasize individual appeals of decisions (L1) and direct participation in systems development (L2). Cities' policy execution departments are not, by their nature, adept at adjusting direction based on external signals.

Furthermore, many cities still approach AI mostly from a pilot project perspective. Attitudes should shift to one of continuous learning and improvement. For example, Amsterdam conducts pilots with uncharacteristically high care. These pilots receive more scrutiny than systems in daily operation to allow for operation "in the wild" while staying within acceptable boundaries. The additional scrutiny throughout and the mandatory intensive evaluations upon completion serve to identify risks that may arise if systems were to transition into daily operation. This careful approach transforms pilots from the non-committal testing grounds common in the business world into something more akin to a social experiment guided by bioethical principles [279]. While Amsterdam's pilots serve as good examples, successful pilots face difficulties in their transition into daily operations. This "innovation gap" (cf. Item T3.1) may be partially alleviated when designers stay involved after delivery. Public AI designers should consider themselves stewards, whose role is never finished [87].

Finally, it is not just AI and its development process that need 'redesigning.' Cities' AI commissioning and governance structures must also be adjusted. Again, referring to our Five Loops Model, this would mean a focus on participation in policy-making (L3) and the second-order feedback loops from decision appeals to developers and policy-makers (L4–5).

Public AI, Urban AI, and VUS

Our example case of camera-car-based trash detection illustrates the need for the public and urban AI fields to converse more actively with each other. Public AI tends to focus on what goes on inside city organizations; urban AI tends to

focus on what happens in the streets. Our results show how the concept of contestability connects the dots between several issues focused on in the literature so far. Namely, between explanations and justifications [48, 86, 106], street-level bureaucrat discretion [109, 300, 301], and citizens' daily lived experience of urban space [219, 220].

Participation in public and urban AI literature is almost invariably of the direct kind [48, 301] as if we have given up on representative modes of democracy. There is potential in renewing existing forms of civic oversight and control. So, again, in our Five Loops Model, a shift from focusing on individual appeals and direct participation in development (L1–2) to participation in policy-making (L3) and monitoring of appeals by policy-makers (L5).

We find it striking that the HCI design space appears to devote little or no attention to (camera-based) VUS. Camera cars appear to offer tremendous seductive appeal to administrators. More public camera car applications will likely find their way into the cities of the global north. They deserve more scrutiny from (critical) HCI scholars.

Speculative Design as a Research Method

Turning to methodological aspects, we will make a few observations. As is often the case with contemporary speculative design, our concept is more a story than a product [119]. Indeed, we sought to spark the imagination of the audience [90, 319]. One respondent recognized this:

> And I think the lack of imagination that you have dealt with really well with your film is what keeps the conversation going even now, which is exactly the goal. (P9)

The story we tell explores the implications of new technology [90]. It is a projection of potential future impacts of public AI that is (or is not) contestable [90]. Nevertheless, it would go too far to say we are 'constructing a public' [74]. We have not engaged in "infrastructuring" or the creation of "agonistic spaces" [110, 119, 151]. We *did* design for one-on-one debate [229] and worked to ensure the video is sufficiently provocative and operates in the emotional register without tipping over into pure fancy or parody [227, 229].

We used speculative design to open up rather than close down [195]. In this opening up, we went one step beyond merely critiquing current public AI practice and offered a speculative solution of contestability, framed in such a way that it invited commentary. Thus, asking questions rather than solving problems may not be the best way to distinguish speculative design from 'affirmative design.' As Malpass [229] points out, rather than lacking function,

critical design's function goes beyond traditional notions of utility, efficiency, and optimization and instead seeks to be relational, contextual, and dynamic.

On a more practical level, by building on the literature [17, 18, 74, 119], we defined success criteria upfront. Before bringing the result to our intended audience, we built an explicit evaluation step into our design process. This step used these same criteria to gain confidence that our artifact would have the effect we sought it to have on our audience. This approach can be an effective way for other design researchers to pair speculative design with empirical work.

4.6.3 Transferability: Results' Relation to City and Citizens

Amsterdam is not a large city in global terms, but populous and dense enough to struggle with "big city issues" common in popular discourse. Amsterdam was an early poster child of the "smart cities" phenomenon. It embraced the narrative of social progress through technological innovation with great enthusiasm. Only later did it become aware and responsive to concerns over the detrimental effects of technology. We expect that Amsterdam's public AI efforts, the purposes technology is put to, and the technologies employed are relatively common.

The city's government structure is typical of local representative democracies globally. Furthermore, the Netherlands' electoral system is known to be effective at ensuring representation. Many of the challenges we identify concerning integrating public AI in local democracy should be transferable to cities with similar regimes.

Amsterdam is quite mature in its policies regarding "digital," including the responsible design, development, and operation of public AI. Less-advanced cities will likely struggle with more foundational issues before many of the challenges we have identified come into focus. For example, Amsterdam has made considerable progress concerning the transparency of its public AI system in the form of an algorithm register, providing explanations of global system behavior. The city has also made notable progress with developing in-house capacity for ML development, enabling it to have more control over public AI projects than cities dependent on private sector contractors.

Amsterdam's residents have a national reputation for being outspoken and skeptical of government. Indeed, city surveys show that a significant and stable share of the population is politically active. Nevertheless, a recent survey shows that few believe they have any real influence.[23] Political engagement and

23. https://onderzoek.amsterdam.nl/publicatie/amsterdamse-burgermonitor-2021

self-efficacy are unequally divided across income and educational attainment groups, and these groups rarely encounter each other.

Our respondents tended to speak broadly about citizens and the city's challenges in ensuring their meaningful participation in public AI developments. However, in articulating strategies for addressing the challenges we have identified, it is vital to keep in mind this variation in political engagement and self-efficacy.

For example, improving citizens' information position so they can participate as equals may be relevant for politically active people but will do little to increase engagement. For that, we should rethink the form of participation itself. Likewise, improving the democratic embedding of public AI systems to increase their legitimacy is only effective if citizens believe they can influence the city government in the first place.

4.6.4 Limitations

Our study is limited by the fact that we only interacted with civil servants and the particular positions these respondents occupy in the municipal organization.

Over half of the civil servants interviewed have a position in the R&D and innovation department of the city. Their direct involvement is mostly with pilot projects, less so with systems in daily operations. The themes and challenges we have constructed appear, for the most part, equally relevant across both classes of systems. It is conceivable, however, that civil servants employed in other parts of the city executive (e.g., social services) are more concerned with challenges we have not captured here.

Further work could expand on our study by including citizen, civil society, and business perspectives. This would bring to the surface the variety of interests stakeholder groups have with regard to contestability measures. Our respondents' statements are based on a first impression of the concept video. We expect more nuanced and richer responses if we give respondents more time to engage with the underlying ideas and apply them to their context. Finally, interviews do not allow for debate between respondents. Another approach would be to put people in dialogue with each other. This would identify how stakeholder group interests in contestability may align or conflict.

4.6.5 Future Work

The public sector context brings with it particular challenges facing the implementation of contestability mechanisms but also unique opportunities. For

example, the existing institutional arrangements for contestation that are typical of representative democracies demand specific forms of integration but offer more robust forms of participation than are typically available in the private sector. For this reason, future work should include the translation of 'generic' contestability design knowledge into context-specific forms. Considering the numerous examples of public AI systems with large-scale and far-reaching consequences already available to us, such work is not without urgency.

Most contestability research focuses on individual appeals (L1 in our Five Loops Model) or participation in the early phases of AI systems development (L2, but limited to requirements definition). Future work should dig into the second-order loops we have identified (L4–5) and how citizens may contest decisions made in later phases of ML development (i.e., L2, but engaging with the 'materiality' of ML [25, 85, 158]). The participatory policy-making loop (L3) is investigated in a more general form in, for example, political science. However, such work likely lacks clear connections to AI systems development implications downstream.

Finally, to contribute to public AI design practice, all of the above should be translated into actionable guidance for practitioners on the ground. Practical design knowledge is often best transmitted through evocative examples. Many more artifacts, like our own concept video, should be created and disseminated among practitioners. HCI design research has a prominent role in assessing such practical design knowledge for efficacy, usability, and desirability.

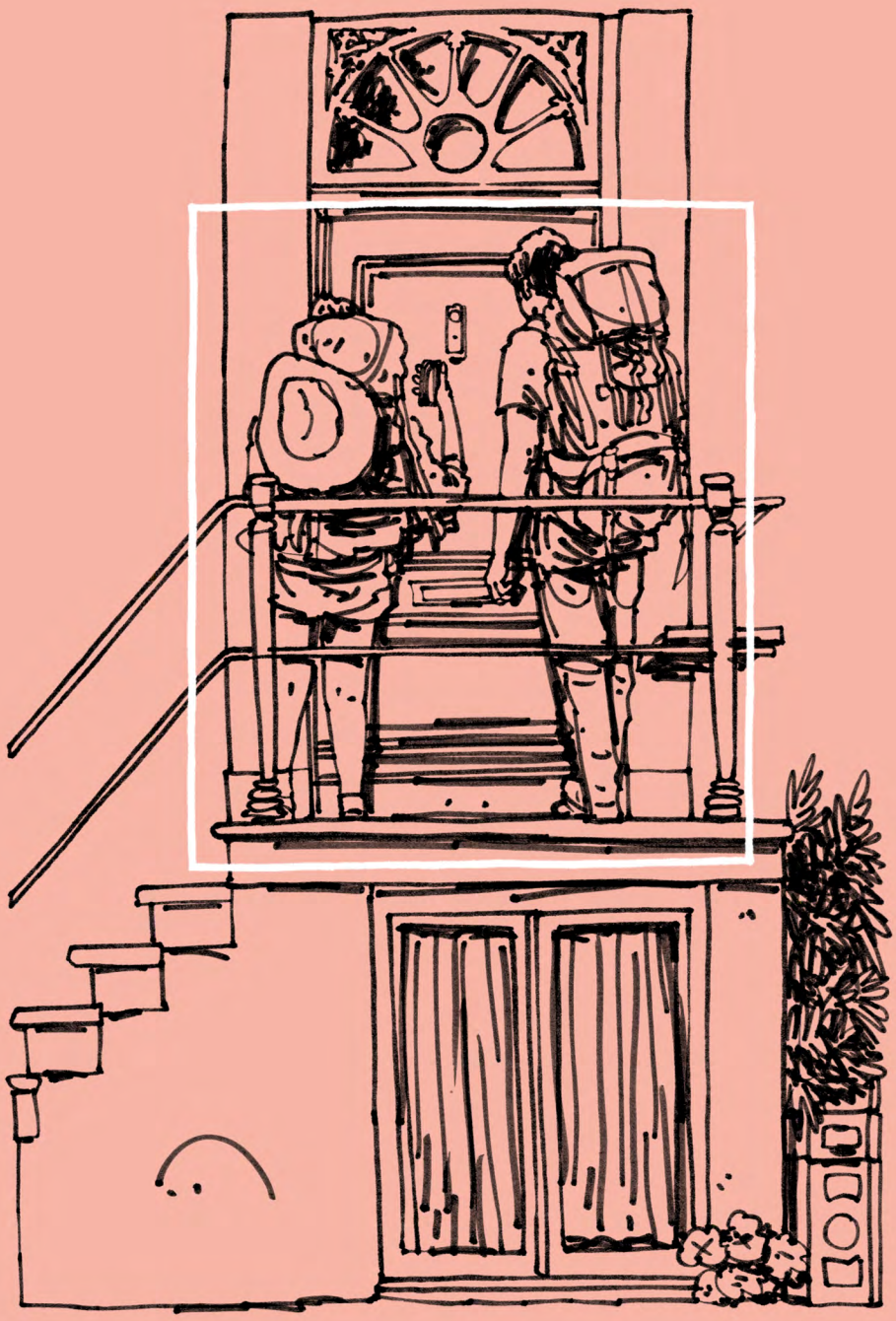
4.7 Conclusion

City governments make increasing use of AI in the delivery of public services. Contestability, making systems open and responsive to dispute, is a way to ensure AI respects human rights to autonomy and dignity. Contestable AI is a growing field, but the knowledge produced so far lacks guidance for the application in specific contexts. To this end, we sought to explore the characteristics of contestable *public* AI and the challenges facing its implementation by creating a speculative concept video of a contestable camera car and conducting semi-structured interviews with civil servants who work with AI in a large northwestern European city. The concept video illustrates how contestability can leverage disagreement for continuous system improvement. The themes we constructed from the interviews show that public AI contestability efforts must contend with limits of direct participation, ensure systems' democratic embedding, and seek to improve organizational capacities.

'Traditional' policy execution is subject to scrutiny from elected representatives, checks from the judiciary and other external oversight bodies, and direct civic participation. The shift to AI-enacted public policy has undermined and weakened these various forms of democratic control. Our findings suggest that contestability in the context of public AI does not mean merely allowing citizens to have more influence over systems' algorithms, models, and datasets. Contestable *public* AI demands interventions in how executive power uses technology to enact policy.

Acknowledgments

We thank Roy Bendor for advising us on our method and approach; Thijs Turèl (Responsible Sensing Lab and AMS Institute) for supporting the production of the concept video; Simon Scheiber (Trim Tab Pictures) for creating the concept video; the interviewed experts for their productive criticism that led to many improvements to the concept video; the interviewed civil servants for taking the time to talk to us and providing valuable insights into current practice; and the reviewers for their constructive comments.



Chapter 5

Envisioning Contestability Loops: Evaluating the Agonistic Arena as a Generative Metaphor for Public AI

Citation:
Alfrink, K., Keller, I., Yurrita, M., Bulygin, D., Kortuem, G., and Doorn, N. "Envisioning Contestability Loops: Evaluating the Agonistic Arena as a Generative Metaphor for Public AI." in: *She Ji: The Journal of Design, Economics, and Innovation* (in press)

Abstract:
Public sector organizations increasingly use artificial intelligence to augment, support, and automate decision-making. Such public AI can infringe on citizens' right to autonomy. Contestability is a system quality that protects against this by ensuring systems are open and responsive to dispute throughout their life cycle. A growing body of work is investigating contestable AI by design. However, little of this knowledge has so far been evaluated with practitioners. To make explicit the guiding ideas underpinning contestable AI research, we construct the generative metaphor of the Agonistic Arena from the political theory of agonistic pluralism. We combine this metaphor and current contestable AI guidelines into an infographic supporting the early-stage concept design of public AI system contestability mechanisms. We evaluate this infographic in five workshops paired with focus groups with a total of 18 practitioners, yielding ten concept designs. Our findings describe mechanisms for contestability proposed by these concept designs. Building on these findings, we subsequently evaluate the efficacy of the Agonistic Arena as a generative metaphor for the design of public AI and identify two competing metaphors at play in this space: the Black Box and the Sovereign.

5.1 Introduction

Algorithmic decision-making in the public sector can undermine autonomy—people's effective capacity for self-governance [239, 283, 292]. To safeguard against this, such *public AI* systems should be *contestable*: open and responsive to dispute throughout their lifecycle, establishing dialogical relationships between decision subjects and system operators [7].

Contestable AI is an emerging field of research within human-centered AI [55]. However, as with other aspects of responsible AI, much of the debate related to contestability has been focused on principles rather than practices [247]. For the contestable AI field's findings to be useable by practitioners, they need to be translated and adapted to specific contexts [135] and presented in ways that they can easily relate to [358]. One such form is *visual explanations*, infographics that represent dynamic processes [337]. Furthermore, design knowledge should be *generative*—allowing for a range of specific solutions without entirely prescribing their form [161]. We can achieve such conceptual richness by articulating a *generative metaphor* [302]—an idea that allows designers to think of a problem in terms of something else, leading to particular diagnoses and accompanying prescriptions.

This contribution hypothesizes a generative metaphor for contestable AI in the public sector context: the *Agonistic Arena*. We construct this metaphor from *agonistic pluralism* [255], a political philosophy that underpins much contestable AI research. Our main aim is to evaluate the Agonistic Arena metaphor's efficacy as a generative metaphor for designing public AI that is more contestable. In support of this aim, we create an infographic of contestable AI that supports practitioners during the concept design of public AI, titled 'Contestability Loops for Public AI.' The infographic builds on previous work, translating contestable AI into more practical guidance [5, 7]. It is also deliberately designed to convey the Agonistic Arena metaphor. We perform a qualitative evaluation of this infographic with practicing designers in a series of workshops. Participants are asked to redesign an existing public AI system to be more contestable, with help from the infographic and the Arena metaphor it embodies.

We frame our approach as constructive design research in the 'field' mode [193, 194]. Our ontological and epistemological commitments are critical realist [115, 132] and contextualist [137, 225]. We use creative design practice to produce an artifact that serves as a research instrument for generating data in a specific context, which is analyzed using interpretative techniques.

This study's contributions are the construction of the Agonistic Arena from political theory, a generative metaphor that animates the contestable AI field

(Section 5.2.4); an infographic that further concretizes and explicates contestable AI knowledge for the audience of design practitioners active in the public AI space (Section 5.3.2); an evaluation of the extent to which practicing designers, when using the Arena metaphor and the Contestability Loops infographic, do indeed produce more contestable concept designs of public AI; (Section 5.5.2) and an account of several *competing metaphors* which may be at play in public AI discourse—the *Black Box* and the *Sovereign* (Section 5.5.3).

This article is structured as follows: First, we provide background on public AI, contestable AI, generative metaphor, agonistic pluralism, and the Agonistic Arena metaphor. Next, we describe our method, including infographic design, workshop focus groups, and reflexive thematic analysis. Subsequently, we describe our results as themes that capture mechanisms put forward by the concept designs created by our workshop participants. Finally, in the discussion, we evaluate the efficacy of the Arena as a generative metaphor for the design of public AI by reflecting on the extent to which the concept design mechanisms are expressions of said metaphor or embody competing metaphors.

5.2 Background

5.2.1 Public and Urban AI

We situate our work in the context of public AI, which we define as the application of adaptive data analysis and processing to enhance, assist, or automate decision-making in the public sector [265, 326].

Research on the use of AI in the public sector, termed *public AI* [48, 86, 102, 106, 109, 230, 277, 300, 301, 346], is growing. While some use "AI" [86, 102, 106], the terms "algorithm" or "algorithmic system" are more prevalent [48, 109, 277, 300, 301, 346]. Such systems inform or automate government decision-making [48, 86, 102, 300, 301]. Key application areas are child protection, public housing, health, social protection, security, and taxation [48, 86, 230]. Main concerns include transparency [48, 86, 106], data collection politics [230, 277], and impact on public sector work [109, 300, 301, 346].

A related field is *urban AI* [68, 220, 221], which delves into AI's role in the built environment. Here, the emphasis is on mobility solutions such as electric vehicle charging, autonomous vehicles, and parking systems [6, 219, 299]. This research examines AI's influence on urban experiences, intertwining AI ethics with urban design ethics [220]. The focus on *spatial justice* [142, 206, 307, 314] is unique to *urban* AI, complementing procedural and distributive justice discus-

sions.

One of the issues relevant to public AI is that of autonomy [239, 283, 292], which the emerging field of contestable AI seeks to address.

5.2.2 Contestable AI

Research on contestable AI has been expanding, highlighting its significance in safeguarding against flawed and unjust automated decision-making by emphasizing human involvement and fostering adversarial discussions between decision subjects and system operators [7, 9, 55, 146, 153, 298, 341].

Contestability can be viewed as humans questioning machine predictions, allowing human intervention to rectify potential machine errors [153, 350]. It can be described as a blend of human and machine decision-making, emphasizing its role in procedural justice and enhancing perceived fairness [222, 341, 366]. The practice of "contestability by design" stresses human intervention both retrospectively and in the AI development processes [9]. Contestability transcends mere human intervention, demanding a dialectical interaction between decision subjects and human controllers [298]. A system's legitimacy is compromised without contestability, which demands *justifications* in addition to explanations [146]. Implementing contestability features in practice will require thoughtful consideration of needs, values, and context [223].

Contestable AI has been conceptualized as systems that are open to human intervention throughout their lifecycle, emphasizing a dialogical relationship with decision subjects. Contestations can be leveraged for continuous system improvement [7]. A proposed design framework lists elements contributing to contestability, incorporating system features and development practices tied to stakeholders and AI system lifecycle stages [7]. Subsequent work emphasizes the relevance of participatory policy-making approaches and the need to monitor contestations for systemic flaws [5].

When we turn principles into practical guidelines, they become more specific but less helpful for orientation. Practitioners might interpret these principles differently than their creators, leading to designs that oppose the original intent. Thus, we use the theory of *generative metaphor* to understand the underlying ideals of contestable AI proponents and convey them more clearly alongside specific principles.

5.2.3 Generative Metaphor

Schön defines *generative metaphor* as a lens influencing our perception and understanding of the world [302]. It involves "meta-pherein," the transfer of perspectives between domains. This perception affects our decisions and actions. For Schön, challenges in social policy stem from problem-framing rather than problem-solving [302]. Recognizing society's implicit generative metaphors enhances our understanding. Not all metaphors are generative; only those offering new insights are. Schön discusses *frame restructuring* to reconcile conflicting perspectives [302].

Related but distinct is Lakoff's *conceptual metaphor*. This theory describes how language uses metaphors to convey deep-rooted concepts, like associating "love" with "warmth." These metaphors connect abstract ideas to familiar sensations, becoming ingrained through cultural interactions [202]. In short, metaphorical thought is unavoidable.

In HCI and design research, generative metaphor has been used to analyze discourses in computing [140, 209] and to analyze user perception of voice interfaces [72]. It has been used to challenge HCI research assumptions [23]. Attempts to formalize the methodical use of metaphor include Method Cards [212], which helpfully categorize them as either weak or strong. Functional prototypes in various domains make use of metaphor for design, including AI [26, 84, 259, 261]. Metaphor and narrative synergistically enhance moral imagination, offering a dynamic approach to the value-sensitive design of AI systems [340]. Metaphorical thinking can foster a more nuanced understanding of artificial intelligence [100]. Metaphor use by designers is inescapable and best done consciously [143].

Generative metaphor reveals that design issues can be interpreted in multiple ways. Each interpretation suggests particular underlying challenges. How we metaphorically frame AI problems matters. Understanding a driving metaphor underpins the effective use of contestable AI prescriptions. The next section delves into this metaphor.

5.2.4 Agonistic Pluralism and the 'Arena' as Generative Metaphor

We see the thinking of contestable AI researchers as shaped by a generative metaphor we call the 'Arena,' taking inspiration from the ancient Greek ideal of democratic competitiveness [105]. This metaphor casts AI systems as a space in which conflict in various forms is embraced and celebrated as a productive

force. *Agonistic pluralism*, the political philosophy underpinning this metaphor was developed by Mouffe [249–255].

Agonistic pluralism presents a democratic model that values productive conflict over deliberation and consensus, emphasizing the celebration of radical differences and contentious expression in democratic practice. It acknowledges the democratic paradox that we can never wholly achieve a thoroughly pluralistic society but argues that conflict is essential to preserving diversity and preventing the erasure of difference. Spaces for contestation must be maintained, allowing dissent and the challenging of power relations. Agonistic pluralism distinguishes between politics and *the political*, focusing on the latter and embracing conflict as intrinsic to societal life. It views diversity of values as constitutive and productive, preventing civic apathy and exposing oppression. In contrast to universal truths, it keeps values open to contestation to promote pluralism and continuous scrutiny of dominant power expressions. Agonistic pluralism sees identities as relational and emphasizes collective identity formation through political participation, opposing deliberative democracy, and aiming to transform antagonisms into legitimate political adversaries engaged in the struggle for hegemony [75, 216, 304].

In the field of science and technology studies (STS), the concept of agonistic pluralism is employed to critique participation and inclusion approaches in responsible research and innovation (RRI). Stilgoe et al. discuss the limitations of inclusion in RRI, suggesting that it often becomes an end in itself, shaped by those in power, and overlooks the diverse motivations of participants [320]. They advocate for more critical reflection on participation and its underlying norms. Van Bouwel and Van Oudheusden argue for a differentiated approach to democratizing scientific governance, pointing out that consensus in democracy often neglects conflict and non-consensual change, advocating for models like agonistic pluralism that embrace disagreement [345]. Genus and Stirling highlight the importance of inclusive, reflexive deliberation in RRI, acknowledging the challenges posed by dogmatism and advocating for incrementalism [126]. Popa et al. focus on the role of conflict in technology history, proposing agonism to manage conflict by valuing responsiveness and dialogue over consensus [281]. Finally, Scott observes that challenges in public engagement in RRI reflect criticisms of deliberative democracy and suggest an 'agonistic' RRI that examines power relations and views stakeholder stances as adversarial rather than equally valid [304].

Researchers have applied agonistic pluralism in the context of interaction design, artificial intelligence (AI), machine learning (ML), and algorithmic decision-making [66, 75, 76, 89, 149, 272, 289]. AI systems seen as objects of agonistic

political design create spaces for confronting power relations [75]. Adversarial design methods can democratize technology development in line with agonistic ideals [272]. The agonistic lens helps us see that AI systems are also always part of contested spaces. When properly agonistic, algorithmic decision-making is always provisional, temporary stabilization of power [66]. Agonistic AI system development would allow society to decide if, when, and how to integrate AI. Agonistic AI decision-making offers the ability for individuals to demand alternative ways of being computed or to reject being computed entirely [148]. Agonistic AI demands broader forms of participation that acknowledge and allow for conflict and are sensitive to power relations and exclusions [289]. Agonism lets us see AI systems not only as a product or producer of politics but also as a space *itself* within which politics happens and to resist simplistic readings of AI's politics as fully liberatory or oppressive [89]. Contra AI safety approaches that rely on principles or technologies, AI development can be conceived of as "machine politics," where agonistic deliberation should not just be the means to achieve AI safety, but its goal [76].

Conceptualizing the Generative Metaphor of the 'Agonistic Arena'

Contestable AI is an expression of the generative metaphor of the Arena. This metaphor casts public AI in terms of a space where interlocutors embrace conflict as productive. Seen through the lens of the Arena, public AI problems stem from a lack of opportunities for adversarial interaction between stakeholders. Prescriptions lean towards making more contentious and open to dispute the norms and procedures that shape AI system design decisions on a global level and human-AI system output decisions on a local level—individual decision outcomes; establishing new dialogical feedback loops between stakeholders that ensure continuous monitoring. The Arena metaphor encourages a design ethos of revisability and reversibility so that AI systems embody the agonistic ideal of contingency.

5.2.5 Design and AI

Although this study's empirical work is centered on early-stage design activities focused on generating concept designs, this should not be taken to mean we hold a linear deterministic view of how design contributes to AI systems. 'Actually existing' AI systems are designed and redesigned on an ongoing basis by groups of people who, more often than not, have a job title that does not include "designer," who do not consider themselves doing design at all, and who are

not necessarily part of the organization designing the system in question. As with other complex sociotechnical systems, (public) AI systems are dynamic, constantly changing in response to feedback from their environment [130].

In this context, design is more akin to what John Seely Brown described as "thinkering" (sic)—experimenting, testing, and adjusting in a collaborative manner akin to the open-source approach [13]. McCullough has described this as 'tuning'— the incremental growth, change, and adaptation of configurations and settings based on the "feel" of the aggregate, something not easily predicted but arrived at iteratively over time based on human judgment [232, pp. 92–94]. Designers become like *stewards* whose role is never finished [87], or *facilitators* of change among a variety of stakeholders, helping them to "act more intelligently" in a more "design-minded way" in the systems we inhabit [330, pp. 7, 214]. As Höök and Löwgren put it, when faced with complex sociotechnical systems that include AI, designers should consider their work as "interventions into ongoing transformations over which they have limited control" [160, p. 34].

Although we evaluate the Arena metaphor in the context of early-stage concept design, its applicability is not intended to be limited to this stage. Instead, we hope it will serve as a guiding concept throughout the AI system lifecycle for all those who contribute to design in some fashion to steer choices towards those that increase AI systems' contestability.

5.3 Method

We aim to develop and evaluate *generative intermediate-level design knowledge* [161, 214], which occupies the middle range between specific instances and general theory, providing seeds for design solutions without prescribing their shape. We build upon prior efforts that introduced a *framework* for contestable AI [5, 7]. Frameworks outline design solution characteristics for achieving goals in a particular context [268]. Our objective is to *evaluate* this knowledge with *practitioners* [134, 135, 167, 364] to strengthen the HCI research-practitioner relationship [135]. We translate the framework and the accompanying generative metaphor of the Arena (Section 5.2.4) into a *visual explanation* [337]. Such infographics are suitable for depicting systems-oriented knowledge and are especially beneficial for practitioners who often rely on visual aids [358]. We conduct workshops with professional designers to assess the infographic, a common method in HCI design research [290, 332]. Our qualitative analysis of workshop outcomes uses the theory of generative metaphor as a lens and utilizes reflexive thematic analysis [43], further adapted using critical realist approaches [118] and annotated portfolios [120, 214].

5.3.1 Preregistration

We preregistered this study at Open Science Framework (OSF).[1] The most notable change between the study plan and this final report is narrowing the focus of the research aim and questions to the efficacy of the generative metaphor of the Agonistic Arena. All data was generated as described, but the analysis scope was narrowed only to cover the generated concept designs.

5.3.2 Visual Explanation Design Process

The process of visual explanation construction was as follows. First, we drafted a creative brief (Appendix B). The two critical ingredients for the infographic are, first, the Contestable AI by Design framework [7], updated with insights from the Five Loops model [5], and second, the Agonistic Arena generative metaphor (Section 5.2.4). The infographic's loops are how the new relations between stakeholders are established, which are an essential element of the Arena. The infographic is specific to the public sector context by explicitly including the representative democratic policy-making process. It is aimed at design practitioners by offering more concrete guidance than the underlying theoretical framework.

Next, we recruited an information designer to lead infographic creation. The primary selection criterion was if their portfolio contained works that resembled the content and style set out in the brief. An innovation lab provided funding for this segment of the study. The infographic went through eight iterations between April 11 and May 22, 2023.

We made some critical design decisions along the way, including the following. A style reminiscent of Chris Ware and his *ligne claire* predecessors (e.g., Hergé, Joost Swarte) creates a legible and relatable look. A2 paper size scale provides sufficient space to include the required detail while still usable on a projected display or printed and kept on the side of a desk while doing concept design work. We included visual references to competition and conflict to strengthen the connection to the Arena metaphor. At a late point in the process, we included a separate element that explicitly describes what motivates contestability: increasing systems' legitimacy over time. Following the pilot workshop on May 10, we made some final adjustments (cf. Section 5.3.3).

1. Under embargo until March 31, 2024, or the time of publication. View-only link to anonymous version: `https://osf.io/qjzgv/?view_only=43f5a7a066cd4e02a9ab3cfb515c877d`.

Visual Explanation

The infographic depicts a generic human-AI decision-making system, four features that create contestability loops, and a fifth section representing the policy and system development context by which a human-AI system is produced (Figure 5.1). The four features are *Interactive Controls*, which allow human controllers and decision subjects to intervene in the AI prediction process; *Intervention Requests*, which enable data subjects to understand individual decisions, express their disagreement, debate system operators, and receive a human review of a decision; *Tools for Scrutiny*, which allow a wide range of groups in society to inspect the workings of human-AI systems; and finally, *Monitoring*, a second-order loop that looks for systemic patterns in individual decision appeals. In the policy and system development part, we show a variety of control means for citizens: electing public representatives, participating directly in policy-making, and participating directly in system development. A separate diagram in the bottom left shows how the human-AI system evolves towards a more legitimate state over time under pressure from repeated contestations.[2]

5.3.3 Design Workshop Focus Groups

We generated the data for this study using workshops with professional designers employed by client services agencies in The Netherlands. In these workshops, we first gave participants a brief introduction to contestable public sector AI and the Agonistic Arena metaphor and explained the infographic. This information mirrors the descriptions in Sections 5.2.2 and 5.3.2. Then, we presented the example case of a real-world human-AI decision-making system piloted in the city of Amsterdam to aid the enforcement of illegal vacation rentals (cf. Section 5.3.3). Subsequently, we asked participants to create concept designs to make this system more contestable. We used the prompt: "Using the infographic for guidance, sketch one or more concept designs to make the vacation rental system more contestable." Participants could work solo, in pairs, or in groups during the design exercise. They were provided with a set of materials to sketch with, which we kept consistent across workshops.[3] We concluded each workshop with a focus group discussion in which participants briefly presented their concept designs. We recorded the audio of these discussions. The first

2. A detailed description of the infographic is provided in Appendix C.
3. A3 marker pad, HB pencils, Sharpie markers, and Post-it notes.

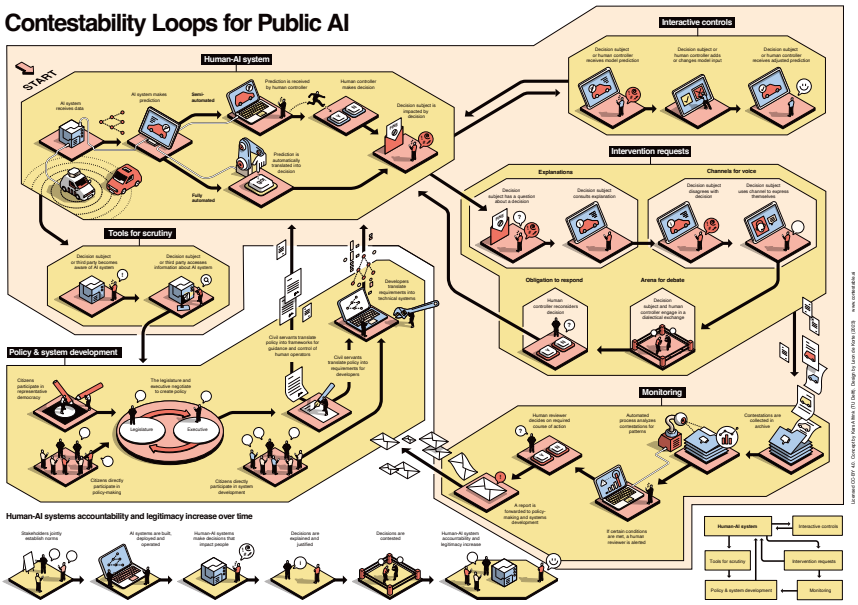


Figure 5.1
Contestability Loops for Public AI infographic used in workshops.

author was the workshop facilitator and lecturer, as well as a guide to the final discussion. We did not actively participate in concept design exercises.[4]

This study received approval from our institute's human research ethics committee. We acquired written informed consent from all participants.

We conducted five workshops at agencies in The Netherlands. Our recruitment strategy was purposive. We sought out interaction design agencies using our network with demonstrable experience with design for the public sector and design for AI or, more generally, data-driven technologies. Participant numbers ranged from three to five (M = 3.6, SD = 0.9). These numbers follow the criteria for focus groups recommended by Braun and Clarke [42, p. 115]. Workshops lasted three hours and took place on participant agencies' premises. Participants spent between 33 and 55 minutes sketching (M = 40, SD = 11). Focus group discussions lasted between 39 and 51 minutes (M = 44, SD = 4). The data

4. The workshop schedule and focus group guide are available in Appendix D and F.

generated consists of concept design sketches and verbal descriptions. Ten concepts were generated in total.[5]

Pilot Workshop

Before data generation, we piloted the workshop with 19 industrial design engineering master students at our institution. Changes we made to the workshop afterward were relatively minor. We included a more detailed walkthrough of the infographic, expanded the case description document with several more example images, and fine-tuned the timing of the various workshop segments.

Participant Demographics

Participants' years of professional design experience ranged from 1 to 35 years (M = 14.3, SD = 10.6). Participants' self-reported knowledge of design for AI ranged from "not at all" to "very knowledgeable" (M = 2.7, SD = 1.0), while their knowledge of design for the public sector ranged from "slightly" to "extremely knowledgeable" (M = 3.7, SD = 1.0).[6]

Case: Illegal Vacation Rental Housing Enforcement Risk Model

For our case, we selected a typical instance of a public AI system. We used the algorithm register of the city of Amsterdam to screen for a system that uses risk scoring because this has become a widespread practice with more than a few public scandals in recent history. We searched for a system that addressed a relatable issue involving some stakes but was not highly polarizing. We opted for a system that the city piloted as part of the enforcement of illegal vacation rentals.[7]

Amsterdam continues to struggle with mass tourism. Visitor levels have rapidly recovered to pre-pandemic levels and continue to increase. Part of the challenge for the city to control visitor flows is the practice of illegal vacation rental properties. The city has two main policy aims. To ensure adequate living space availability for residents and to prevent visitors from adversely affecting the city's livability.

5. Concept designs are summarized in Appendix G.
6. On a scale of one to five; one being "not at all" and five "extremely knowledgeable."
7. The full case description document provided to participants is included in Appendix E.

In early 2020, the city announced a pilot system that would aid in screening reports of possible illegal vacation rentals. The system would help the city save time on finding suspicious homes, freeing up time for investigating properties.

The system takes as input reports from citizens about possible housing fraud. The system then selects additional data available on the property. The probability of housing fraud is calculated by the system using a model created using random forest regression and historical data on investigated reports. The system uses SHapley Additive exPlanations (SHAP) [218] to calculate the contribution of features to the prediction. Based on the report, risk score, and explanation, a civil servant decides whether or not to investigate. Surveillance and enforcement officers conduct the investigation and submit their findings to an enforcement lawyer. The enforcement lawyer decides if there is a violation or not.

Issues include high fines that can lead to undesirable situations where enforcement is deemed disproportionate to the violation, such as an honest mistake. As designed, the system lacks contestability.[8]

5.3.4 Analysis

Our overall analysis approach is based on reflexive thematic analysis [39–41, 43–45]. We adapt the approach to our purposes, drawing inspiration from critical realist approaches to TA [118, 359]—in particular, alternating between data-led and theory-led coding, as well as a hierarchy of codes and themes that reflects our research question (Figure 5.2). We took further inspiration from the annotated portfolios approach to design knowledge construction from individual design instances [35, 120, 214].

Data analysis was performed by the first author. The remaining authors contributed with partial coding and review of coding results (cf. Section 5.3.5).

Data Preparation

To prepare data, we scanned sketches and stored them as image files. Focus group audio recordings were first machine-transcribed using Whisper.[9] The first author then manually edited the raw transcriptions, removed identifying

8. This system was never fully piloted due to the pandemic and the introduction of new legislation—notably the requirement of a permit and registration number—which made other forms of enforcement that do not depend on reports—but make use of scraping vacation rental websites—more feasible. See council information letter on results of housing fraud enforcement (May 23, 2023): `https://amsterdam.raadsinformatie.nl/document/12800007/1`
9. `https://github.com/openai/whisper`

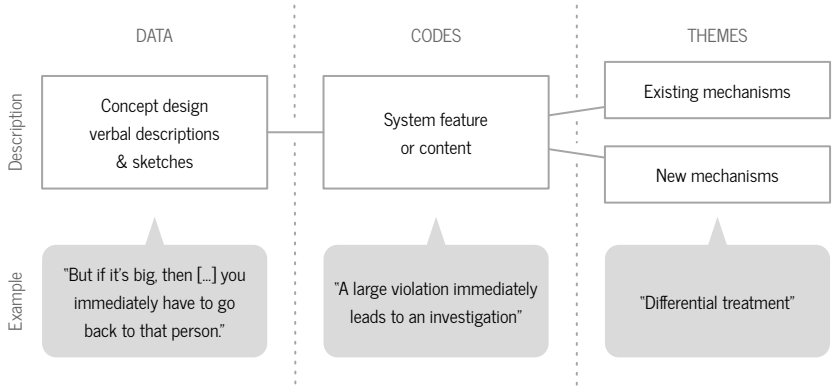


Figure 5.2
Conceptual model of thematic analysis of verbal concept design descriptions and accompanying sketches. The data we work with are transcripts of verbal descriptions of concept designs supported by sketches. We code the transcripts for verbal statements that refer to system features or contents. These codes are grouped into higher-level themes that each represent a single mechanism for contestability. These mechanisms are compared to the infographic to determine whether they count as *new* or *existing*.

details, and added speaker identification pseudonyms (e.g., "P1"). For those focus groups conducted in Dutch (workshops 2 and 5), the transcripts were subsequently translated into English using Google Translate and manually edited by the first author. We stored each concept design description in a separate text file. The remainder of the focus group discussion was not the subject of the analysis reported on here.[10]

Thematic Analysis

The first author coded transcripts in Atlas.ti following the conceptual model outlined in Figure 5.2. We first coded the transcript on the sentence level for statements describing system functionality or contents. Next, we standardized and consolidated codes using consistent language and theoretical concepts. We then organized codes into themes, each representing a *mechanism*: a discrete process or technique that enables contestability. We discarded codes that did not fit this scheme. Finally, we compared each theme to the features described

10. Data is archived and made available on 4TU.ResearchData: `https://doi.org/10.4121/8eb71eb5-cc7f-4055-aba3-2e90812a940b`.

by the infographic. Those mechanisms that resembled infographic features we considered *existing*. Mechanisms that did not resemble the infographic we considered *new*. We referred to the concept design sketches throughout this process to contextualize the analysis.

5.3.5 Credibility Strategies

To improve the credibility of our analysis, we had discussions among team members to ensure a more thorough analysis. By using reflexivity, we accounted for our particular positions and how these might affect our analysis. Peer debriefing with colleague researchers was an external check on our research process. Member checking—sharing a draft report with participants for feedback—ensured our analysis reflects participants' intentions.

5.3.6 Positionality

We favor contestability and would like contestable AI to be taken up in practice. The participants are peers in the design field, some of whom we have previously worked with. They are employed by design agencies, some of whom we have professional relationships with. The case is from the city of Amsterdam, a municipality we have worked with on other studies in the past.

5.4 Results: Concept Design Mechanisms

Participants generated a total of ten concept designs.[11] Concept descriptions are summarized in Table 5.1. Figure 5.3 shows examples of concept design sketches produced by participants. From these designs, we construct existing and new mechanisms. We summarize these results in Tables 5.2 and 5.3. For each concept design, we indicate the absence or presence of each mechanism. We further distinguish between partial and full presence. We assign partial presence for concept design descriptions that contain a mere one to two references to the mechanism, usually on the level of a coherent utterance.[12]

11. Sketches and summary descriptions of these concepts are available in Appendix G.
12. Concept designs are referred to with a C followed by a number (e.g., 'C2' is the concept generated in workshop two). If a workshop produced more than one concept, these are given a suffix (e.g., 'C1.1' is the first concept generated during workshop one). Participants are referred to with a W and a number to indicate the workshop they were part of, followed by a P and a number to indicate the workshop's individual participant (e.g., 'W1P1' is participant one in workshop one).

Table 5.1
Summaries of concept designs.

ID	Summary
1.1	A transparent and equitable system for monitoring citizens' behavior in Amsterdam, focusing on detecting illegal renting practices, with annual assessments, anonymous reporting, and an open algorithm, complemented by a non-intimidating AI character for communication and guidance.
1.2	A visible indicator system for properties rented out on platforms like Airbnb, enhancing complaint handling and neighborhood impact awareness through data integration and company involvement in rental distribution.
1.3	A system for equitably sharing unused space, focusing on positive reinforcement and contextual analysis to pair individuals with a feedback loop for shared financial gains and a nuanced approach to handling infractions.
2	A system that gathers data and provides decision subjects, like landlords or affected individuals, with transparent, disputable reports and visual representations of decision-making factors, emphasizing the need to mitigate biases at both AI and interpretation levels for fair and unbiased outcomes.
3	An open, collaborative system prioritizing transparency, dialogue, and feedback, focusing on providing comprehensive information, engaging users and developers, ensuring a human approach in decision-making, and continuously improving fairness and effectiveness with public and law enforcement input.
4.1	A two-dashboard system aimed at combating fraud and enhancing transparency, with one dashboard offering individual case insights and the other providing policymakers and the public with aggregated data on fraud trends, contributing factors, and bias monitoring.
4.2	Focuses on enhancing transparency and fairness in handling fraud reports by making algorithmic processes understandable and contestable to citizens and experts while addressing challenges like bias and policy implications.
4.3	A process that encourages empathy and understanding by allowing for the contestation of legislation, reports, and algorithmic analysis, aiming to improve fairness and effectiveness through collaboration between the accuser and the accused.
5.1	A system for Airbnb that identifies and assists vulnerable hosts who unintentionally commit fraud, offering a transparent, step-by-step resolution process with opportunities for feedback and intervention by an enforcement officer.
5.2	A circular, transparent system for handling potential fraud, combining data analysis with SHAP explanations, human judgment, and communication to validate reports, assess fraud likelihood, and decide on proportionate actions while minimizing administrative burdens.

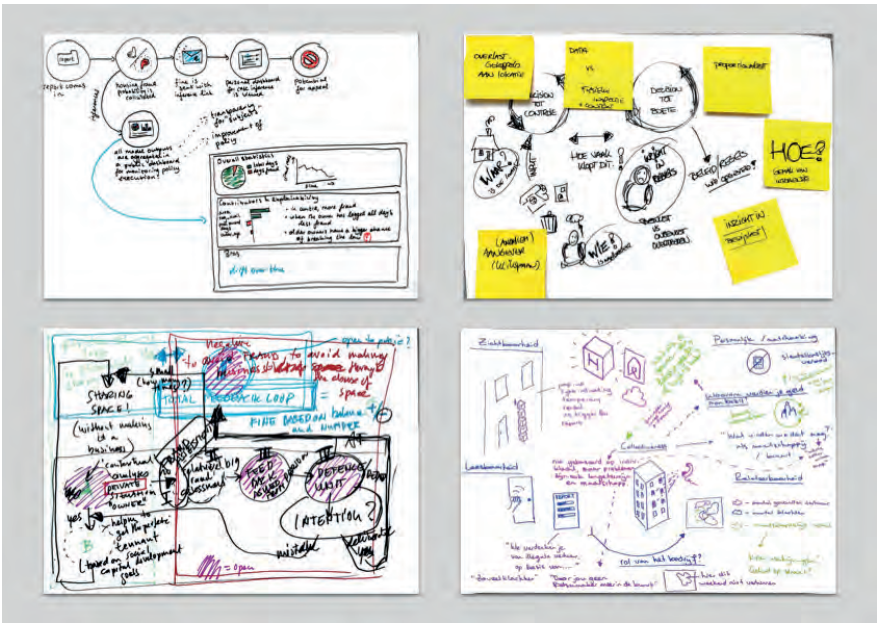


Figure 5.3
Examples of concept design sketches created by participants during workshops.

5.4.1 Existing Mechanisms

The existing mechanisms that feature most prominently are Explanations, Interactive Controls, Intervention Requests, and Tools for Scrutiny.

Explanations

Explanations can be delivered through a variety of offline and online touchpoints. When an inspector visits a subject, they should bring a report explaining the reason for the investigation (C1.2). Explanations should seek to reduce subjects' emotional pressure from being under investigation (C3). Some concepts explicitly suggest the use of visual communication (C2).

In terms of contents, explanations should include details of the report (C3), the data collected on a subject (C2, C3), and the reasons for the risk score (C3). Explanations should also include the reasons for being investigated (C3) and details of the decision-making procedure (C2). Explanations show how a subject's group characteristics may impact their risk score and downstream treat-

Table 5.2
Occurrence of *existing* mechanisms in concept designs.

	Concept design									
Mechanism	1.1	1.2	1.3	2	3	4.1	4.2	4.3	5.1	5.2
Explanations	○	◐	○	●	●	●	◐	○	○	◐
Interactive controls	○	○	○	◐	○	○	●	○	●	●
Intervention requests	○	○	●	◐	◐	●	●	○	◐	●
Monitoring	○	○	○	○	◐	○	○	○	○	◐
Participatory policy-making	○	○	○	○	◐	○	◐	○	○	○
Participatory system development	○	○	○	◐	●	○	◐	○	○	○
Tools for scrutiny	◐	◐	○	○	●	●	●	○	◐	◐

Legend: ● present; ◐ partially present; ○ absent.

ment (C3). Ideally, explanations match the information inspectors use to decide to investigate (C2).

An explanation is also included in case of a fine (C4.1). These contain the details of the perceived violation and related regulations (C5.2). They also again show all the data that went into the decision (C4.1), and they should clearly state how to pay for a fine (C4.2). Finally, explanations are a starting point for contestation (cf. Intervention Requests) (C2, C4.1).

> But once you get that charge, explaining it is really important because right now, on most websites, when you are charged for something it's not, I can't understand what the charge is. What exactly is that charge? How has it been levied? (W4P2)

Interactive Controls

Controllers need to understand the AI system because they use their outputs. The global-level explanations towards this end can be technical but not too much (C4.2). Enforcement officers (human controllers) have discretion. They are the ones who decide to visit a reported residence. To exercise this discretion, they need to receive an explanation of why it has been flagged (C5.1).

In the pilot system, this explanation is provided using SHAP. The system should also show the confidence of the prediction (C5.2). When they review predictions, controllers should also be able to adjust them. A controller should be able to provide qualitative feedback on a prediction. Such feedback and the reviewing controller should be recorded for future reference. If, at a later point,

a subject is fined, the original prediction, along with the controller's review and feedback, should be reproducible (C4.2).

Decision subjects should be able to correct data collected about them if it is incorrect and respond to the submitted reports (C2, C5.2). To this end, they should be notified when a report has been submitted, and the AI system has produced a risk score about them (C5.2). They should be able to respond to the reports themselves (C5.2). Subjects could also have access to an 'open desk' where they can speak to a civil servant, receive an explanation, inspect, and possibly adjust input data (C5.1).

> It starts a bit with the reports that are there, of nuisance, and so on. I also thought of making that clear to the subject. Whether he also thinks that those reports are justified or correct or at least knows about them. (W5P3)

During a visit, the enforcement officer completes a checklist. The decision subject can inspect this report as well. If the subject indicates they disagree with the decision that the officer arrives at, an objection procedure can be started right away (C5.1, C5.2).

Intervention Requests

Several concepts aim to increase the agency of decision subjects (C2). Subjects should be made aware of the fact that contesting is possible and that it is allowed. The system should explain the appeal procedure. Contesting should be easy and require the minimum administrative hassle (C5.2). Several concepts propose some form of notification to alert subjects of a decision to fine and the possibility of contesting (C3, C4.1, C4.2). Such a notification may lead to a personal dashboard on a website, which explains the indicators that went into the decision, and a way to contest the various aspects of a decision or to satisfy the fine is made available (C4.1, C4.2).

Others propose an 'open desk' as the touchpoint for requesting human intervention and initiating an objection procedure (C5.1). A subject's defense will help determine if they made a mistake or a deliberate violation (C1.3). There might be time limits on these contestations (C4.2). When system operators review a decision in response to an intervention request, the subject receives feedback. This feedback again includes a justification for the ultimate decision and any outstanding action items for the subject (C5.1). Even so, it is preferable not to incorrectly fine people in the first instance instead of enabling them to correct mistakes afterward. Even if human intervention is easy to acquire, correcting mistakes requires much effort (C5.2).

> Then, you can have the convenience of intervention, but Jill only wanted a week's vacation. That was just, yeah... and then suddenly you are in a paper tiger, and you spend a year trying to prove that you live on 1A and not on 1B. (W1P4)

Tools for Scrutiny

Some participants see an agonistic relationship between the government and the public intrinsic to system development. However, communication should emphasize that it is not about government versus citizens. It should emphasize that confrontation is a form of dialogue and is considered a positive (C3).

> I think that's also key indeed in that way of communication that it's not really about us versus them or like this government versus public thing. Simply because of how the system is made, there are two intrinsic kinds of perspectives to things, but if there is openness for both of the parties to improve the system, I think it's good. (W3P2)

Citizens should grasp how the global system functions (C1.1, C4.2). While technical details benefit experts, they can be confusing for others. The goal is to simplify the system for widespread understanding (C3).

Some recommend that platforms like Airbnb display local regulations when users create city listings. These platforms should also explain enforcement methods, including the role of the AI system (C4.2). Others advocate for a "softer approach," emphasizing the reasons behind regulations and the use of AI. The city must highlight the system's societal benefits and purpose (C3).

A proposed solution is a public monitoring dashboard for both the public and policymakers. This dashboard would present aggregate data, including number of days that fraud was detected, decision-impacting features, and bias measures like model drift over time (C4.1, C4.2). Recognizing bias might lead to feature adjustments (C4.2). Another suggestion involves monitoring the two phases of decision-making: investigation and fining. Developers can analyze these phases for error rates, with each requiring different human judgment (C5.2).

One concept suggests a website that breaks down the AI system. It would explain the AI's role in decisions, the data used, and its impact on outcomes (C4.2). Another concept includes publicizing data from decision-appeal monitoring (C3).

Lastly, one concept suggests placing signs outside vacation rentals to increase community awareness (C1.2).

> And then [...] give people, like, a light to hang out on the outside of their house to indicate whether it's a hotel room for a night. Like the New York hotel signs. And then maybe you could make an Airbnb one or a Booking one, just make it visible so that you know, like, there's a lot of noise there, and it's actually

> currently rented out. So I know where my complaint will go. I don't think it's a practical solution, but I like it anyway. (W1P2)

5.4.2 New Mechanisms

The most prominent new mechanisms include Annual Assessments, Differential Treatment, Input Data Revisions, Pro-Active Notifications, and Pro-Social Behavior Incentives.

Table 5.3
Occurrence of *new* mechanisms in concept designs.

	Concept design									
Mechanism	1.1	1.2	1.3	2	3	4.1	4.2	4.3	5.1	5.2
Annual assessments	●	○	○	○	●	○	○	○	○	○
Differential treatment	◐	○	●	○	○	○	○	○	◐	●
Input data revisions	◐	◐	○	●	●	○	○	◐	○	○
Pro-active notifications	◐	○	○	○	◐	○	○	●	●	○
Pro-social behavior incentives	○	●	●	○	○	○	○	◐	○	○

Legend: ● present; ◐ partially present; ○ absent.

Annual Assessments

All citizens receive an 'annual assessment,' which includes the data collected on them, a provisional risk score, justifications of the current policy, opinions of the various political parties on this policy, savings on civil servant labor, and related performance indicators. When the city introduces the system, everyone starts with a clean slate. One's 'status' is also periodically reset (C1.1, C3).

> But this is the rule we have now, and you have violated it clearly. You have rented out for 40 days, ten days in excess, and I have to pay. And you know this because in your annual check, this is the file, so at least you know. I really do think a lot of people don't even know. (W1P1)

Differential Treatment

Several concepts include a measure for varying the penalty for a violation based on the scale of the violation, its nature, or the subject's circumstances (C1.3, C5.1, C5.2). Such variability would open up space for negotiation between subjects and operators. Fining should consider subjects' knowledge, understanding, and

intentions (C5.2). The system should weigh the costs of an infringement against the social benefits a person is delivering by renting out their home (C1.3). Similarly, enforcement should be justified in a legal sense *and* a "human" one (C5.2). Although small 'mistakes' may be tolerated initially, they could add up and lead to scrutiny from enforcement as well (C1.3). The enforcement officer should make this distinction (C5.2). Monitoring of decision appeals should also look for indications that enforcement is not proportional to the scale of the violations.

> So it's more like, I think, more related to justice [...] or how can you make it a system. (W1P3)

Input Data Revisions

The system is an example of so-called reports-driven enforcement augmented with AI. Several concepts addressed the perceived limitations of these reports as input data (C1.1, C1.2, C2, C3). Asking citizens to report on each other can be problematic. Citizens can abuse the system to report on others they conflict with (C3). Furthermore, the channel used for reporting can influence data quality (C3). Civil servants should screen reports before recording them if reporting happens via phone or some other synchronous medium. This screening should also apply to people who report others. The system should include the identity of the person submitting the report in the subsequent risk assessment of the residence (C1.1, C3). The number of people reporting on the same residence should also be a factor in the risk assessment (C1.1, C3).

A couple of concepts suggest pulling in additional data to mitigate the limitations of these reports (C1.1, C3). Further downstream, the controllers who evaluate the reports with the AI system's aid can also be biased. One concept proposes specific measures against this (C2). The system should inform the human controller that the reports and accompanying input data can also be biased (C2). Finally, one concept addresses that reporting citizens need to properly understand how and by whom their reports are processed (C1.2).

> Right, so I was first thinking you filed a complaint, but you don't know what the effect of that complaint is. So you don't know whether it will go to, like, I don't know, I used to have an alcoholic neighbor. So maybe it will go to, like, a social system or to the Airbnb system. (W2P2)

One concept anticipates that some reports originate from disputes between the reporting person and the reported citizen. This concept proposes creating a framework for resolving such disputes without the city acting as a direct intermediary (C4.3).

Pro-Social Behavior Incentives

Several concepts address the social issue of vacation rental fraud and the negative impact of mass tourism more directly. They consider the presence of an algorithmic system for enforcement an opportunity to encourage more pro-social forms of vacation rentals.

Because vacation rentals have collective impacts, but individual complaints drive the enforcement policy, another concept seeks to help Airbnb hosts 'see' the impact on their community by pulling in more data related to such impacts and visualizing it alongside the rental platform interfaces. The aim is to 'nudge' users to refrain from renting if there is too much pressure on a neighborhood (C1.2).

Conversely, some concepts acknowledge that vacation rentals can also be socially desirable. For example, they can lead to new social connections or allow for the use of living space that would otherwise remain unoccupied. Negative consequences happen when people turn vacation rentals into profit-seeking businesses. These concepts seek to encourage such pro-social forms of vacation rentals (C1.2, C1.3).

To curb the adverse effects of profit-seeking vacation rentals, a couple of concepts suggest redistributing the gains of rentals within communities (C1.2, C1.3).

> And then I got into this kind of path of thought that it's, this is all, it's all based on individual incidents. I think the effects of Airbnb are collective as well, so it changes neighborhoods and not just noise levels… during one night. So, for example, like, there are fewer supermarkets and more bike rentals, and this kind of systemic impact. But now it's just based on individual complaints and individual cases. And I think there should be more indicators than just individual complaints. (W1P2)

Pro-Active Notifications

Several concepts include measures to ensure subjects are actively made aware of critical events in the systems' process, including being reported, being flagged for investigation, and the availability of an objection procedure (C5.1, C4.3, C3, C1.1).

When someone reports a subject, they receive a notification with a preview of the algorithmic assessment. This preview can also be a starting point for a subject's contestation of the report or the system's assessment. (C4.3). This same notification should also include a means of making reparations. The person who filed the report can then indicate satisfactory reparations, in which case

the case is dropped (C4.3). The notification of being reported should not identify the reporting person.

Further downstream in the process, when a controller has opted to investigate a residence, the subject should again be alerted. This notification should again include an explanation of the decision. The explanation should also include instructions on contesting the decision (C5.1, C1.1, C3).

Several concepts include convenient touchpoints for indicating disagreement in the real world, for example, when an inspector visits. When a subject formally does so, the system should initiate an objection procedure and notify the subject when it has become available for them to act on (C5.1).

> But yes, that's right. I think we wanted all the decisions that were made, whether you got into such a box at all, back as quickly as possible or as easily as possible to the person involved. Who might want to fight it. (W5P1)

5.5 Discussion

5.5.1 Summary of Findings

We analyzed ten concept designs and constructed from them mechanisms that were either already present in our infographic (*existing mechanisms*) or were not, and therefore considered *new mechanisms*. The mechanisms are summarized in Table 5.4. Their relationship to the example case human-AI system and the three generative metaphors we have constructed are shown in Figure 5.4.

Table 5.4
Summary of metaphors, mechanisms, and concepts. *New mechanisms are italicized.*

	Mechanism	Description	Concept designs
Agonistic Arena	*Differential Treatment*	Varying penalties based on the nature of the transgression or a subject's circumstances so that enforcement becomes more proportional.	C1.1, C1.3, C5.1, C5.2
	Input Data Revisions	Accounting for the inherently biased nature of reports, mostly by including additional data to contextualize reports.	C1.1, C1.2, C2, C3, C4.3

Table 5.4
Summary of metaphors, mechanisms, and concepts (continued).

	Mechanism	Description	Concept designs
	Interactive Controls	Provisions for human controllers to review, adjust, and provide feedback on risk scores. Means for citizens to respond to report contents and correct input data.	C2, C4.2, C5.1, C5.2
	Intervention Requests	Provisions for subjects to inspect and contest sanctions, mostly through websites or physical touchpoints.	C2, C4.2, C5.1, C5.2
Black Box	Explanations	Describing the data and procedures that lead to a penalty, delivered through personalized websites or face-to-face interactions with street-level bureaucrats who perform home inspections.	C1.2, C2, C3, C4.1, C4.2, C5.2
	Pro-Active Notifications	Ensuring that a subject is made aware of the fact that they are under scrutiny at every step of the process.	C1.1, C3, C4.3, C5.1
	Tools for Scrutiny	Integration of AI system details into rental platforms. Public monitoring web-based dashboards with a variety of aggregated performance metrics.	C1.1, C1.2, C3, C4.1, C4.2, C5.1, C5.2
Sovereign	*Annual Assessments*	Risk scoring all citizens every year and proactively informing them of their profile should a report be filed.	C1.1, C3

Table 5.4
Summary of metaphors, mechanisms, and concepts (continued).

Mechanism	Description	Concept designs
Pro-Social Behavior Incentives	Leveraging the AI system to transform the underlying social issue, e.g., by mediating between reporters and renters or raising community awareness about the measured social cost of vacation rentals.	C1.2, C1.3

We will now proceed to answer our main research question: *What is the efficacy of the Agonistic Arena as a generative metaphor for the design of public AI?*

5.5.2 Public AI as Agonistic Arena: Beyond Agreeing to Disagree

The Agonistic Arena frames public AI as a space where all forms of struggle are celebrated as productive. It finds expression as practices that seek to establish new discursive relations between stakeholders, enable continuous monitoring in the interest of contingency and admittance of fallibility, and create sociotechnical arrangements prioritizing mutability and reversibility.

Existing mechanisms that express the Agonistic Arena are Interactive Controls and Intervention Requests. New mechanisms that do the same are Differential Treatment and Input Data Revisions.

Interactive Controls enable civil servant discretion, a necessary component of anticipatory flexibility at the level of individual decisions. Human controller-initiated adjustments of inferences are also an implicit signal useful for monitoring. Interactive Controls enable citizens to ask an alternative calculation of themselves. *Intervention Requests* is a necessary component of any contestable AI system, so it is no surprise that almost all concept designs include some form. It enables the contestation of individual decisions. Some concept designs devote more attention to the discursive element, preventing appeals from becoming a mere one-way expression of discontent and not a rearrangement of power relations. *Differential Treatment* is related to street-level human-AI discretion and ensures more proportional algorithmic enforcement. It enables a diversity of possible algorithmic decision outcomes and, as such, can make systems more pluralistic and inclusive. *Input Data Revisions* makes subject to contestation the

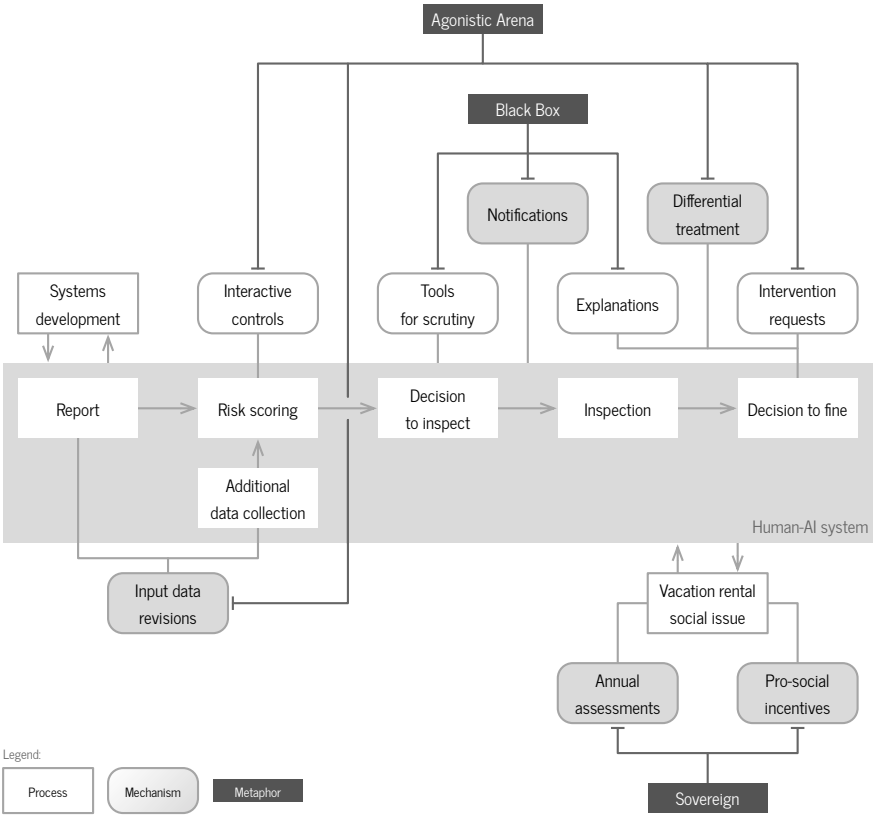


Figure 5.4
Diagram summarizing findings. The example case of the human-AI system process (square boxes with white fill) is related to the *existing* and *new* mechanisms proposed by the concept designs (rounded corners, white and grey fill, respectively), which in turn are related to the three generative metaphors (dark grey fill).

data that serves as input for inferences and acknowledges the contingent social nature from which the data, reports in particular, originate. It establishes new relations between reporters and subjects and allows for mutability.

We can see that two out of four existing mechanisms and two out of five new mechanisms can be construed as expressions of the Arena, indicating that our participants thought of public AI in those terms. This ratio suggests that the Arena is a suitably generative metaphor that reframes AI's problem in terms

aligned with agonistic pluralism: a need for more discursive relationality, contestability, and contingency.

However, we consider the remaining mechanisms to *not* be expressions of the Arena. One may expect that existing mechanisms align with agonistic pluralism's priorities because they match the infographic's elements. However, a closer examination of how the concept designs concretely instantiate these mechanisms suggests otherwise. Less surprising, perhaps, is that more than half of the *new* mechanisms are expressions of a metaphor other than the Arena.

Next, we will describe two candidates for what these alternative framings, these competing generative metaphors, might be. We arrived at these metaphors using abductive reasoning [213]. They are our best assumptions for the metaphors that design workshop participants may have used. The metaphors are primarily based on workshop findings, contextualized by our familiarity with contemporary AI design ethics and political discourse. In the future, further work could be done to confirm the recurrence of these metaphors across a broader range of design settings.

5.5.3 The Black Box and the Sovereign—Two Competing Metaphors

Existing mechanisms that *do not* express the Agonistic Arena, but a competing metaphor are Explanations and Tools for Scrutiny. New mechanisms that do the same are Annual Assessments, Pro-Social Behavior Incentives, and Pro-Active Notifications. We see two competing metaphors in the design space covered by the concept designs: the Black Box—AI as an opaque system that requires explanation—and the Sovereign—AI as a benevolent overseer to which social coordination can be delegated.

The Black Box: Sunlight Is the Best Disinfectant

The Black Box is a prominent competing metaphor in our participants' concept designs and public thought about AI in general. The Black Box focuses on the presumed opacity of AI systems, i.e., a lack of transparency, and that they require *explanations* to be trustworthy and accountable [267]. This opacity can stem from secrecy, illiteracy, or scale and complexity [51]. The Black Box metaphor is central to the field of explainable AI (XAI) [2, 20], which develops technical solutions to the fundamental opacity of ML models.

The existing mechanisms that express the Black Box metaphor are Explanations and Tools for Scrutiny. The *new* mechanism that does the same is Pro-Active Notifications.

Explanations describe the technical process factors that lead to a decision. The explanations proposed by most concept designs are insufficient for contestability because they lack the normative dimension, i.e., they do not offer a "justification" [146] of why a decision is desirable. Without justification, a decision subject cannot mount an "articulate act of defense" [298]. As instantiated by the concept designs, Explanations align with the liberal conception of deliberative democracy, where facts and reason alone are sufficient to make a case. *Tools for Scrutiny* seeks to make AI more transparent and explain its workings globally. Its implementation is limited to a fact-based, technical account in most concept designs. Our participants' conception of this mechanism does not embrace any particular computation's contingent and contested nature. It typically leaves out or underemphasizes the importance of including the norms governing AI systems' functioning. Finally, the *Pro-Active Notifications* mechanism lacks the two-way dialogical nature necessary for true contestability. It is also unclear how the tempo of the procedures that subjects receive notifications about intersects with their ability to halt procedures before the next stage commences. In this way, notifications reinforce the top-down authoritarian nature of the system rather than destabilize it.

The distinction between the Black Box and Arena metaphors in contestable AI literature emphasizes transparency and accountability, suggesting a shift from merely factual to normative explanations. It argues for replacing opaque models with interpretable ones, particularly in high-stakes situations [293], enabling operators to exercise discretion in applying decision rules [8, 34, 235, 276]. Additionally, it proposes a sociotechnical approach, focusing on collective understanding and dialogue [11], rather than individual interpretation, to overcome limitations in fully explaining machine outputs [354]. This approach critiques the Black Box for its neglect of power dynamics and its unrealistic assumption of a liberal ideal of free and equal individuals in practice.

The second and final competing metaphor to discuss, the Sovereign, very much acknowledges power. However, rather than distributing it downwards to citizens, it pushes it *upwards* to a machinic autocrat.

The Sovereign: Save Us From Ourselves

A less prominent but intriguing metaphor we dub the Sovereign is expressed by the Annual Assessments and Pro-Social Behavior Incentives mechanisms, both new.

The metaphor of the Sovereign frames social problems as stemming from a lack of coordination toward common interests. Under this view, society's problems are too complex for individuals to comprehend the repercussions of their actions. Therefore, what is needed is an all-knowing, all-seeing, all-powerful, but benevolent machine to which people delegate this coordination. Individuals willingly give up their freedoms and accept the imposition of this Sovereign on their daily lives in return for the peace of mind that whatever the AI asks them to do will contribute to the common good. This common good has been decided upon beforehand and encoded in the AI overseer.

The mechanism of *Annual Assessments* assumes a future in which the system preemptively processes all citizens periodically and makes risk scores continuously available. It is autocratic because computation is inescapable. The system imposes a single worldview through calculation. At the same time, it is paternalistic because it considers preemptive calculation a positive, which helps citizens adjust their behavior to avoid sanction. Politics is still possible in this vision, as citizens are informed about parties' views of the current calculative regime, which citizens can presumably consider at the next election. However, politics has been purged from the realm of policy execution entirely. In a sense, it is the logic of New Public Management—the total separation of policy-making and policy execution—taken to its very extreme [88, 365]. *Pro-Social Behavior Incentives* uses data collection and processing to visualize impacts to discourage harmful vacation rentals or, conversely, to incentivize pro-social forms through various data-driven credit schemes. Coordinating social actions is removed from the local sphere and delegated upwards to an autocratic data-driven apparatus.

Ironically, the issue of authoritarian AI is a longstanding concern in critical AI studies. Typically, algorithms are perceived as computational tools, making authoritative choices between variables to deliver a single result. This becomes problematic when we aim to influence their decision-making processes [66, pp. 79, 86–87]. Concerns over the imposition of data-driven cybernetic choice architectures are also enduring in critical smart cities research [111, 185, 233, 295, 308].

No matter how enlightened and benevolent, the AI as Sovereign is fundamentally at odds with conceptions of public AI as an Arena.

5.5.4 Relationships between Arena, Black Box, and Sovereign

Here, we briefly examine the relationships between the three metaphors, drawing on Hochuli et al.'s framework of politics, post-politics, and antipolitics [155, p. 13].

The Black Box metaphor represents a postpolitical stance, suggesting that providing more information could resolve public AI issues without considering AI's inherently political nature. In contrast, the Arena metaphor demands not just explanations but also justifications, advocating for the empowerment of individuals to hold AI system operators accountable. This approach aligns with a political perspective, emphasizing a return to active politics.

The Sovereign metaphor differs significantly, aligning with antipolitical currents. It proposes an authoritarian solution to the complexities of democratic deliberation, placing decision-making authority in a "machinic" leader rather than a human one.

In essence, current public AI aligns with the technocratic post-politics of recent decades. The Black Box metaphor, although acknowledging the lack of accountability in this system, fails to envision a clear alternative and leans towards a neoliberal worldview. The Sovereign metaphor critiques this order's inadequacies and suggests eliminating politics altogether, ironically with AI's assistance. The Arena metaphor acknowledges similar frustrations but advocates for further democratization of AI and emphasizes political contestation [258].

5.5.5 Implications for Design

We see two competing metaphors at play—Black Box and Sovereign—that allow designers to think of public AI in terms other than that of an Arena, and as a result, these metaphors pull concept generation in another direction.

Implicit generative metaphors shape our thinking. When designing, or indeed when we are communicating design knowledge, it is helpful to be explicit about our own. Moreover, if we seek to change the way designers frame problems related to AI, crafting new metaphors will be necessary. Such metaphors can be assembled from theory, as is the case in concept-driven interaction design research [321], such as we did here, with our appropriation of the political philosophy of agonistic pluralism for the drawing out of the generative metaphor of the Agonistic Arena.

Designers can view the infographic and its associated contestable AI framework differently from what its creators intended. Clarifying the central metaphor can help designers better understand tools and techniques. However, tool creators can never transfer intent entirely reliably. For example, some of our participants adapted tools for contestability more in line with the Black Box metaphor, which is popular in HCI design. This mixup could be because our participant designers favor a fact-based, consensus-seeking democratic view.

While some participants may support the anti-authoritarian notion of contestable AI, they also proposed ideas echoing the Sovereign. This paradox highlights how political beliefs can sometimes be contradictory [176]. Even designers focusing on human-centered AI in public settings can hold conflicting views. We risk creating inconsistent proposals without thoroughly understanding the political philosophies influencing our designs and articulating a coherent stance.

5.6 Conclusion

Contestability is a quality that ensures public AI systems respect people's autonomy. The emerging field of contestable AI has developed principles and practices. However, designers require more contextual guidance and productive concepts to consider public AI aligned with contestable AI ideals. To this end, we constructed the generative metaphor of the Agonistic Arena from works that apply the political theory of agonistic pluralism to design and AI. We then created a visual explanation illustrating various system features that increase public AI systems' contestability. This infographic makes explicit visual reference to the Arena metaphor. We evaluated the infographic with practitioners in a series of design workshops. We analyzed the resulting concept designs for their shared mechanisms. We distinguished between mechanisms that are already present in the infographic and those that can be considered new. We reflect on these mechanisms in light of the Arena metaphor to show that four out of nine mechanisms can be traced back to it. The remaining mechanisms we can interpret as stemming from competing metaphors.

Since metaphorical thought is inescapable [202], and since using particular metaphors to frame design challenges leads to particular diagnoses and accompanying prescriptions [302], design research and practice are well-served by the explicit development and deployment of metaphor. Our findings show how the theory of generative metaphor can be used in constructive and analytic ways to evaluate intermediate-level design knowledge. At least three metaphors occupy the public AI design space—the Arena, the Black Box, and the Sovereign. If we aim to ensure public AI systems respect autonomy through contestability, the Sovereign should be opposed, the Black Box should be considered insufficient, and we should embrace the Arena.

Acknowledgments

We thank Responsible Sensing Lab for funding the infographic production and general support; Thijs Turèl for all his support; Rob Collins (Umeå University) for suggesting the Arena metaphor; Leon de Korte for creating the infographic; Stijn van der Meulen for help with making sense of the vacation rental example case; all students who joined the pilot workshop; all designer participants for their time, effort, creativity, and engagement; and Sem Nouws for suggesting the Black Box metaphor.

Chapter 6
Discussion and Conclusion

In this concluding chapter, I first revisit the aim. After this, the research questions are answered in sequence. Next, I reflect on the implications of the findings for research, society, and design. Subsequently, some of the limitations of this work are addressed. I conclude with some suggestions for future work.

6.1 Recalling the Aim

This thesis aims to address iniquities related to autonomy by exploring the design of public AI systems for *contestability*, which is defined as being open and responsive to civic dispute. In Chapter 2, together with my co-authors, I explored tensions between experts' and citizens' conceptions of AI transparency, using field tests of a high-fidelity prototype for a transparent smart EV charge point. In Chapter 3, using a systematic literature review, we constructed a provisional design framework for contestable AI. In Chapter 4, we mapped out challenges facing the implementation of contestability in public AI using a speculative concept video and semi-structured interviews with civil servants. Finally, in Chapter 5, we formulated three competing generative metaphors for public AI using an infographic and focus group workshops with interaction design practitioners.

6.2 Answering the Research Questions

I will now review the research questions and answer each in turn.

6.2.1 RQ1: What are the diverging conceptions of transparency between experts who design, develop, and govern public AI systems and citizens who use those same systems?

Recall that the question being addressed in this context was a critique of the feasibility of full transparency in sociotechnical systems due to ethical and practical limitations and the difficulty in fully comprehending such systems. I concluded that explanations—how AI transparency is achieved—should be tailored to specific audiences and contexts. The objective was to gain insight into how experts and citizens perceive transparency. This understanding, in turn, allows us to improve information provision and control.

From the results, my co-authors and I constructed three tensions that characterize citizens' and experts' conceptions of transparency: (1) information quality vs. quantity; (2) AI opacity vs. other matters of concern; and (3) informational needs vs. control.

Experts believe transparency is achieved by providing truthful information about automated decisions. They expect citizens to be able to assess system fairness using this information and that it is actionable. Citizens are largely indifferent to AI unless it mediates other matters of concern. They primarily experience explanations as burdensome and question their relevance if not accompanied by the ability to override system decisions.

So, transparency should not be seen as a property of technology but as a communicative process between experts and citizens, *mediated* by AI. Understanding a system is not the product of simply receiving and processing information. Understanding emerges from the debate between stakeholders and is always provisional. All three tensions point to the need for additional channels for voice through which this debate can be facilitated.

6.2.2 RQ2: What socio-technical features and practices contribute to AI system contestability?

If accountability and legitimacy are our goals, then transparency alone is insufficient. Contestability is proposed as a viable alternative where transparency and explainability cannot or should not be achieved. It is how subjects can demand justifications from system controllers. Because most of the existing work on contestability, as so much else on the ethics of AI, is focused on general principles, my co-authors and I sought to map out practical guidance in the form of those 'active ingredients' that contribute to the contestability of AI systems.

The findings consist of a framework describing five system features and six development practices contributing to AI system contestability. They are mapped to typical human-AI system actors and system development lifecycle phases. The features and practices are summarized in Tables 6.1 and 6.2.

The framework takes a sociotechnical perspective by focusing many of its recommendations on AI systems' entangled and volatile nature. Interactive Control fosters a continuous negotiation process between artificial and human agents. Explanations encompass the entire behavior of automated decision-making systems rather than focusing solely on the technical models. These are complemented by Intervention Requests, which facilitate a dialectical interaction between decision subjects and human controllers that is closely linked with artificial agents. Additionally, the framework describes Tools for Scrutiny that document the technical aspects of AI systems and the processes involved in their construction. This emphasis on transparency extends to ex-ante safeguards, which certify entire organizations rather than isolated technical systems. Another component is Agonistic Design, which reveals how values are embedded within specific sociotechnical arrangements. This practice creates platforms for stakeholders to collaboratively shape decision-making processes, reflecting a diversity of perspectives and interests. The framework also addresses the volatility of AI systems through Quality Assurance (QA) practices, both during and after development. QA During Development involves iterative building and testing, possibly in a living lab setting, to ensure the system's adaptability and responsiveness. Post-Deployment QA focuses on maintaining traceable decision chains involving human and artificial agents, ensuring accountability and transparency. Finally, the framework emphasizes the importance of Risk Mitigation, which involves educating human agents on responsible and effective ways to interact with AI systems. This educational aspect is crucial for fostering an informed, conscientious approach to AI.

The framework emphasizes embracing controversy and disagreement in system design to drive continuous improvement. Instead of seeking consensus upfront, it suggests setting up procedural mechanisms to manage disputes. Stakeholders need not agree on every aspect of the system's design or decision outputs but must concur on procedures for handling disagreements. A potential risk is that designers exploit these procedures to be negligent. The procedures should require accountability for decisions leading to disputes, ensuring transparency in decision-making to mitigate this risk. The adversarial approach embodied by the framework prioritizes democratic values like inclusion, plurality, and justice over mere efficiency.

Table 6.1
Features that contribute to contestability.

Feature	Examples
Built-In Safeguards	External adversarial system; formal constraints.
Interactive Controls	Negotiate, correct, or override machine decisions; feedback loop back to training; supplement local contextual data.
Explanations	Traceable decision chains; behavioral explanations; sandboxing; local approximations; justifications.
Intervention Requests	Human review; supportive, synchronous channels; third-party representation; collective action; dialectical exchange.
Tools for Scrutiny	Norms linked to implementation; documentation; formal proofs; comparative measures; opaque assurances.

Table 6.2
Practices that contribute to contestability.

Practice	Examples
Ex-Ante Safeguards	Anticipating impacts; acceptance criteria; certification.
Agonistic Dev Approaches	Co-construct decision-making process; ongoing adversarial dialogue.
QA Measures During Dev	Stakeholder needs guiding development; bias prevention; living labs; stakeholder feedback.
QA Measures After Deploy	Procedural integrity; monitoring for bias & misuse; feedback from corrections, appeals & additional contextual info.
Risk Mitigation	User education; environmental limits.
Third-Party Oversight	Model-centric tools for auditing; trusted intermediaries; secure environments.

6.2.3 RQ3: What are the challenges facing the implementation of contestability measures in public AI?

For this question, I focus more narrowly on the use of AI in the public sector. Although there is limited research in this area, a growing body of work emphasizes the importance of involving impacted communities in the development of solutions, understanding the discretion of frontline civil servants, and establishing communication between citizens, civil servants, and developers. These factors are all essential elements of contestability. Thus, our goal was to understand better the challenges that arise during the implementation of contestability in public sector practice.

The findings show that public AI contestability efforts must contend with limits of direct participation, ensure systems' democratic embedding and seek to improve organizational capacities.

Enabling civic participation: Citizens need skills and knowledge to contest public AI on equal footing. Channels must be established for citizens to engage city representatives in a dialogue about public AI system outcomes. The feedback loop from citizens back to system development teams must be closed. The city must mitigate against reporting inequality and the limitations of direct citizen participation in AI system development.

Ensuring democratic embedding: Public AI systems are embedded in various levels of laws and regulations. An adequate response to contestation may require policy change before technology alterations. Oversight by city council members must be expanded to include scrutiny of the executive's use of AI. Alternative non-legal dispute resolution approaches that integrate tightly with technical systems should be developed to complement existing complaint, objection, and appeal procedures.

Building capacity for responsibility: City organizations' fragmented and bureaucratic nature fights against adequately responding to citizen signals. More mechanisms for accountability are needed, including logging system actions and monitoring model performance. Civil servants need more knowledge and understanding of AI to engage with citizens adequately. New roles that translate policy into technology must be created, and more multidisciplinary teams are required. Contracts and agreements with external development parties must include responsible AI requirements and provisions for adjusting course mid-project. Contestability requires time and money investments across its various enabling components.

Contestability in the context of public AI does not mean merely allowing citizens to have more influence over systems' algorithms, models, and datasets. *Contestable public AI demands interventions in how executive power uses technology to enact policy.*

6.2.4 RQ4: What strategic guiding concept best complements contestable AI prescriptions of a more tactical nature?

The reasoning behind the fourth question was based on the fact that the political philosophy of agonistic pluralism is a prominent theme in AI thought, whether explicitly or implicitly. In addition to the tactical guidance already covered in our provisional design framework, designers require a guiding concept to think about public AI in agonistic terms. With this in mind, we aimed to develop a generative metaphor for contestable AI from theory and assess it with design practitioners.

The findings illustrate how the design of contestable AI can be guided by the metaphor of the Agonistic Arena, which casts public AI systems as spaces where (1) conflict is embraced by making matters subject to contestation; (2) discursive relations between a plurality of stakeholders are created; and (3) fallibility is acknowledged by ensuring flexibility and continuous monitoring for issues.

At least three metaphors occupy the public AI design space—the Arena, as mentioned earlier, the Black Box, and the Sovereign. The Black Box casts the problem of public AI as a lack of transparency, "sunlight is the best disinfectant." The Sovereign invites us to see AI as a means of solving social coordination problems, "save us from ourselves." If we aim to ensure public AI systems respect autonomy through contestability, *the Sovereign should be opposed, the Black Box should be considered insufficient, and we should embrace the Arena.*

6.2.5 Overarching Research Question: What socio-technical design interventions enhance the contestability of public AI systems?

Finally, I turn to the overarching research question. In this thesis, I aimed to explore the potential harm to human autonomy caused by public AI systems. To address this issue, I proposed the principle of human control of technology, which, seen through an agonistic lens, led to contestability as a desirable system quality. My objective was to find sociotechnical design interventions contributing to this quality. Having answered the four research questions, I can now formulate an answer to the overarching research question.

Transparency is essential for contestability, but alone, it is not enough. To increase the contestability of systems locally and globally, we need to implement system features and development practices that promote it, both before and after the fact. Moreover, we should continuously monitor individual contestations to identify systemic group-level issues. Citizen participation is valuable but not the only way to achieve contestability. We should also rely on representative forms. In addition to system properties, we can generate additional interventions by thinking about public AI systems as if they are Agonistic Arenas, which helps to enhance the contestability of specific public AI system instances in their respective contexts.

6.3 Implications

In this section, I discuss the implications of the findings for the philosophical account of autonomy in relation to AI, research into responsible, explainable,

and human-centered AI, participation in and contestation of public AI systems, and design research and practice.

6.3.1 Autonomy Seen Through a Contestability Lens

So, how do the findings add to, fit with, extend, or challenge the account of autonomy with which I motivated this research?

Autonomy can be understood as *effective self-governance* and can be seen as having an authenticity and an agency dimension [283]. Contestability supports autonomy in the face of AI systems primarily by enabling *human control*. Fjeld et al. identify measures for human control that include explanations, notifications, means of appeal, human review, and opt-out procedures [108]. Furthermore, subjects should be able to intervene in the development of systems prior to implementation. My co-authors and I also identified these measures through a literature review of contestable AI (Chapter 3). They are also all present in concept designs produced by participants working with the Arena metaphor and Contestability Loops infographic (Chapter 5).

Rubel et al. argue that autonomy demands that AI systems that people are subjected to should be ones they can reasonably endorse [292, pp. 45–69]. They describe information and means of control that we owe individuals if we are to respect their autonomy [292, pp. 70–96]. Finally, they describe how subjects can legitimate AI systems only if their autonomy is sufficiently ensured [292, pp. 163–183]. I compare this account with my findings in the remainder of this section.

Reasonable endorsement: If people cannot reasonably endorse a system they are subjected to, they would likely want to contest it somehow. Furthermore, a subject can be considered within their rights to contest a system when it is unreliable, makes use of inputs a person cannot or should not be held responsible for, has decision outcomes that involve significant stakes, or distributes its outcome burdens unequally across groups. In a sense, these aspects could be seen as a template for the justifications advocated for by Henin and Le Métayer [145], which are part of our framework's feature of Explanations (Section 3.6.1).

When building their case, a decision subject or group of subjects could use the reasonable endorsement test to structure their argument that forms the basis for an Intervention Request. This idea recalls the "scaffolding for learning" that Vaccaro et al. [342] advocates, which is also part of our framework's feature of Intervention Requests (Section 3.6.1). From a design perspective, we could even scaffold such argument-building through appeal interfaces and design deeply integrated alternative dispute resolution procedures along similar lines.

The reasonable endorsement view is compatible with contestable AI. It adds further detail to what grounds a subject may have for contestation.

Informed agency and informational control: The Explanations feature in the framework can provide information for practical and cognitive agency. The principles of informed practical and cognitive agency and informational control are useful guides for determining what should be included in explanations (cf. Section 3.6.1). Practical agency demands that explanations enable subjects to correct or mitigate system effects—usually technical and operational information. Cognitive agency demands enough information to exercise evaluative control over one's life. Informational control, the ability to correct information fed into a system, aligns with Interactive Controls in the framework (Section 3.6.1) and Intervention Requests (Section 3.6.1). This account of informational provisions and control is compatible with the findings on transparency and explainability. By viewing information as a resource for exercising agency, we bring clarity to the aims of transparency and explanations.

Legitimation and democratic agency: Justifications are the resource subjects can use to legitimate a system. There are two pathways to legitimation: normative authority and democratic will. Both depend on subjects' autonomy, and contestability is the quality that ensures mechanisms are in place for subjects to exercise it.

In the case of an appeal to normative authority, contestability mechanisms, chiefly Tools for Scrutiny (Section 3.6.1), are how subjects and their representatives gain access to the information that will enable them to evaluate its justifiability. In this way, contestability satisfies the access constraint.

In the case of democratic will, the contestability-by-design practices identified in the framework, primarily the Agonistic Development Approaches (Section 3.6.2) but also, to some extent, the various QA Measures (Sections 3.6.2 and 3.6.2), are how system development is democratized. The autonomy account adds the normative constraints that should bind this democratic pathway. I rarely see this detail sufficiently emphasized in works exploring participatory AI system development. An exception would be Himmelreich [152], who similarly argues that only a "thick" conception of democracy—one that not only allows for participation but also includes deliberation over justifications—will address some of the current shortcomings of AI development. My concern is that, as it currently stands, participatory AI development practices run the risk of arbitrariness.

To sum up, reasonable endorsement can be seen as providing the grounds for contestations. Informational needs for agency recast explanations as tools for exercising control. The legitimation pathways of normative authority and

democratic will further elucidate the value of contestability for a democratic society.

6.3.2 Responsible, Explainable, and Human-Centered AI

Several implications can be drawn out that are relevant to responsible, explainable, and human-centered AI. I frame them as things, based on the findings, citizens can demand from experts.

Trustworthiness: Experts should not think they can improve trust by simply 'explaining things better.' Trust requires the possibility of citizen control. Explanations focused on justifying systems are more likely to create trust. The same applies to opportunities for voice. Above all, what produces trust are systems that operate predictably and reliably. No amount of transparency will compensate for poor system design. In case of system failure, contestability is the mechanism by which subjects can ensure recovery and repair.

Matters of concern: Rather than addressing AI opacity as an isolated issue, explanations should focus on how AI mediates other issues that people may care about. This mediation view means the connection with policy aims and governing norms must be made more explicitly so that those can shape the accounts offered to citizens downstream.

Conversely, those translating policy into human-AI systems for execution must become more attuned to potential and actual conflicts of interest that may arise from AI mediations. System developers should become more adept and comfortable immersing themselves in such controversies.

Information quality: Citizens should demand better quality explanations rather than simply more information. This focus on quality includes the contents of the explanation and how it is delivered. Explanation modalities should be designed to match the contexts and activities within which people encounter them. This importance of context means that the design of explanations must become more user-centered and participatory.

Responsibilization: Citizens should reject being made responsible for assessing AI systems. It should be reasonable for citizens to expect public AI systems to meet adequate safety and reliability norms. However, in the case of failure, it should be trivial to report issues, and deploying organizations should be prepared to respond to such signals efficiently and effectively. This required capacity for responsivity and flexibility has consequences for design and engi-

neering choices. Technologies that make failures costly or impede affordable change should be avoided.[1]

6.3.3 Participation and Contestation

In this section, I draw out several implications for participatory and contestable public AI governance, design, and development.

Quality assurance versus compliance: Currently, cities' procedures for handling citizens' reports, questions, complaints, objections, and appeals (an escalating ladder) are often decoupled from primary processes and are usually handled by separate departments. The rationale appears to be that these procedures are in place for legal and governance compliance and not much else. This disconnect impedes information flow between citizens, civil servants who interact with them (e.g., call center workers), and system developers. As a result, feedback loops that may serve as a means for quality improvement are suboptimal or nonexistent. Cities should work to leverage contestations for quality assurance purposes, not just legal compliance.

Deep integration: Appeal procedures benefit from 'deep integration' with primary processes. For example, Vaccaro et al. argue true contestability operates "in band" of systems—using a metaphor taken from telecommunications which refers to the practice of sending control information over the same channel as the one used for the primary data stream [341]. This deep integration allows iterative human-AI decision-making that is impossible in the more common 'out of band' scenario. Such a requirement could also function as a break on unrestrained uses of AI. Suppose the costs of implementing sufficiently integrated contestability do not outweigh the ostensible benefits offered by the automation of a process. In that case, deployers should reconsider the implementation of AI in the first place.

Dialogical relations: Cities should increase the possibilities for dialogical relationships in the middle range of the report-question-complaint-object-appeal spectrum. The fieldwork and literature studies indicate that a sizeable amount of effort and attention is devoted to ensuring possibilities for dialogue and debate at the far ends of this spectrum. It is usually possible for a citizen to talk to a human with a reasonably low barrier to entry when reporting or asking a question. On the other end of the spectrum, in cases of proper appeals, citizens usually find themselves in the legal system where the hearing of both sides is

1. This recalls insights from Collingridge [64] and the field of Responsible Research and Innovation (RRI) more broadly [126].

established practice. The middle range, the transition point between these two ends, lacks accessible means of dialogue. Worse, the erosion of legal protections and the reversal of the burden of proof sometimes accompanies the automation of policy execution.[2] This *juridification* of relations is not only detrimental to citizens' well-being, but it is also ultimately costly for deploying organizations. Alternative forms of dispute resolution (e.g., forms of mediation), possibly taking inspiration from peace studies [294], could keep citizens and governments out of costly legal procedures and serve as the basis for the feedback loops and deep integration previously mentioned.

Democratizing the democratic control of AI: Cities should revitalize the representative democratic mechanisms for indirect citizen participation in policy development and its downstream system development. Much of participatory AI research is premised on a private sector context (i.e., interactions with big tech platforms) or spaces where representative democratic rule of law has been significantly eroded and captured (e.g., the Anglo-American context), or researcher preferences that are suspicious of institutions and lean towards direct democracy. As a result, prescriptions tend to ignore the opportunities afforded by existing representative democratic institutions in which public AI systems are embedded.

Although it can offer much, as this research has shown, direct forms of civic participation run up against significant limits (e.g., lack of inclusion). Powerful actors can easily coopt them. An alternative avenue that is, in my view, promising to pursue is the revitalization of representative democratic oversight and control on a local, national, regional, and global level. In the Dutch case and regimes like it, this requires significant investment in elected politicians' knowledge, time, and resources for taking the initiative regarding how policy governs and is translated into AI systems. In addition, beyond electoralism, a fruitful avenue of exploration would be forms of people's councils and other radical democratic forms of institution building [216, 233, 256].

6.3.4 Design Research and Practice

The implications for design research and design practice are discussed in tandem here because much of this work has operated at its boundary.

Designers as stewards: Local governments' innovation pipelines rely heavily on pilots. In my fieldwork, I saw that Amsterdam pilots can be conceived of as social

2. The 2023 scandal around the use of a biased fraud detection system by the Dutch Ministry of Education executive agency serves as a particularly stark example of this dynamic. See, for example, the report by Helwig [144].

experiments guided by bioethical principles [279], instead of the non-committal testing grounds common in the business world. This practice is something to aspire to, but such pilots still do not address the so-called 'innovation gap'—the struggles even successful pilots face to be implemented in daily operational practice. One cause for this is the common practice of insisting on a complete handover of a piloted system from innovation to executive departments. To address this, new design roles should be created that act as stewards, guiding projects from pilot settings into production and beyond [87]. This idea resonates with Hill's suggestion to create strategic design roles inside governments in the service of social innovation [150].

Speculative design as research method: It is common practice to distinguish speculative design from traditional 'affirmative' design by saying the former practice asks questions while the latter solves problems [119]. From a design research perspective, this is not a very useful distinction. In this thesis, we have employed speculative design as a research method to ask questions. However, the answers people responded with were the data that was analyzed for knowledge generation in service of problem-solving. I prefer Malpass's framing of speculative design as having a function that goes beyond traditional notions of utility, efficiency, and optimization and instead seeks to be relational, contextual, and dynamic (i.e., "para-functionality") [229].

Multi-activity design knowledge: For design practice, research could do more to couple low-level practical or tactical guidance with higher-level strategic or conceptual guidance. My intuition is that combining these two rough categories of levels of abstraction in service of the same design aim is more effective than one or the other in isolation.

The work presented in Chapter 5 was motivated by observing how design practitioners would apply the low-level guidance in the framework and create design solutions that did not feel quite right. It became clear that this was mainly due to a conceptual drift from contestability to a focus on transparency. This perceived drift was the impetus to develop a means of articulating and communicating the higher-level guiding concept that could accompany the more low-level prescriptions.

The use of generative metaphor [302] or the related notion of framing [79] for conceptual guidance in design is not new. Our contribution is mainly to call out the value of coupling such guidance—typically aimed at the level of strategy or vision development—with that of a more tactical nature. The strong concepts proposed by Höök and Löwgren [161] resemble this proposal. However, they approach the issue differently, seeking to embody only part of a complete design solution while pointing to use practices. Where strong concepts seek to offer

guidance of a strategic and tactical nature through the same vehicle, we suggest doing the same through two distinct knowledge communication artifacts that are aligned and closely coupled.

Another way of making the same point departs from a conception that design knowledge can contribute to one of three types of design activity (Figure 6.1): ideation (i.e., framing and reframing), specification (or planning), and form-giving (i.e., aesthetics, how a design appears to the human senses).[3] Researchers creating intermediate-level knowledge should seek to address all three activities, perhaps not always all of them in equal measure. However, they ignore any one of these at their peril.

Visual communication of design knowledge: Surprisingly, little work investigates the efficacy of particular forms of presenting design knowledge. A welcome exception is formed by the work of Gray and Kou [e.g., 135]. The infographic was generally well-received by participants, particularly for its usefulness, learnability, and flexibility. From my own experience, I can also say with some confidence that designers make copious use of a broad range of usually off-the-cuff information designs to make sense of design problems and more polished ones to convey their diagnoses and prescriptions to clients and other stakeholders. However, the participants were less readily able to reflect on their preferred presentation forms of knowledge they 'consume' as part of their work. I feel a space could be opened for investigating the visual design *of* design knowledge presentation forms.

6.4 Limitations

This section acknowledges several limitations of this thesis' approach related to context, participants, and data generation.

6.4.1 Context

The framework and the broader research focus on the algorithmic decision-making context, mainly in the public sector—as opposed to other AI systems (Chapter 2). The design work was specific to a few particular types of public and urban AI applications: smart EV charging context (Chapter 2), camera cars (Chapter 4), and risk models (Chapter 5).

3. This conception is primarily adapted from Löwgren and Stolterman [215] who propose a trifold design process model that encompasses "vision," "specification," and "operative image."

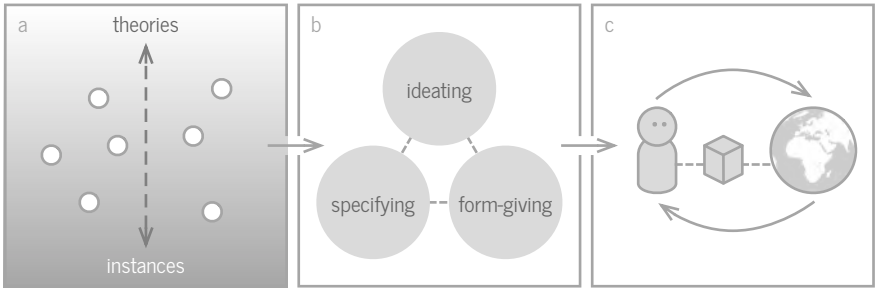


Figure 6.1
Conceptual framework of how design research and practice relate, adapted from Höök and Löwgren [161], Löwgren and Stolterman [215], and Verbeek [347]. Design research knowledge occupies a middle range between general theory and specific instances (a). Knowledge contributions can serve the design activities of ideating, form-giving, and specifying (b). Design activities are directed at intervening in present and future human-thing-world mediations (c).

The smart EV charging system (Chapter 2) is, on a software level, deterministic instead of probabilistic, as in the case of ML. Despite this, from a sociotechnical perspective, it is still sufficiently complex and unpredictable for concerns over autonomy and control to be relevant.

Almost all empirical work was specific to the Amsterdam context, with its own particular local representative democracy, reasonably mature digital policy, and engaged and self-efficacious citizenry (Chapters 2 & 4).

Future work could branch out to other AI applications outside public administration and urban governance, different technology stacks, and locales not in the Global North's developed core.

6.4.2 Participants

The participants tended to be from well-off and highly-educated strata of society (Chapters 2, 4 & 5). Most civil servants I worked with are from the innovation arm of the city government (Chapter 4). The design practitioner participants were all from design agencies in The Netherlands (Chapter 5).

Future work should involve stakeholders from other socio-economic classes. A challenge is to go beyond actors directly implicated in systems of study. For example, citizen groups who are indirectly impacted could be equally relevant to study but are much harder to reach out to. Similarly, it can be reasonably straightforward to involve civil servants who sit in the middle tier of organizations, particularly when they have innovation as part of their remit. It is far

more challenging to involve executive decision-makers on the one hand and frontline workers on the other.

6.4.3 Data Generation Methods

The provisional framework is based on a small sample of primarily theoretical papers. Most source papers lack application context (Chapter 3). However, since its publication, contestability studies have notably increased from high-level conceptual concerns to lower-level practical ones.

Given a PhD project's time and resource constraints, the data generation methods typically operated on short timescales. Longer timescales of engagement could be attempted. Given the lifecycle view of AI systems development, empirical work following concerns related to contestability throughout an entire cycle would be very valuable. Similarly, the design activities I studied were limited to concept design only (Chapter 5). These activities could be expanded to additional design phases and implementation and post-implementation stages—i.e., "design after design" [92].

Other formats other than 1:1 interviews would allow for more debate between stakeholders (Chapter 4). To some degree, this was already born out during the focus groups reported on in Chapter 5. However, those participants were colleagues who generally shared a similar outlook. Research that aspires to agonistic democracy would do well to seek out controversy rather than avoid it.

6.5 Future Work

Future work on contestable AI could take several directions. Here, I outline three auspicious, urgent, and novel ones.

Explainability, discretion, and everyday life: Researchers could develop a program that connects public and urban AI concerns that are typically dealt with separately, namely: (1) explanations and justifications; (2) street-level bureaucrat discretion; and (3) citizens' daily lived experience situated in urban space. Such a program could map out information flows and decision points across these networks. I expect that the discursive relations between citizens and bureaucrats demand alternative forms of AI transparency and explanations. I also expect a better understanding of discretion will lead to human-AI decision-making arrangements more attuned to people's needs.

Participatory policy development and ML engineering: Researchers could develop approaches to participatory system development that engage with the 'materiality' of ML. Much of participatory ML thus far deals with the requirements setting

of development. I expect such an approach will demand new methods and tools for hands-on ML collaboration between developers and citizens. Researchers could also develop approaches to participatory policy-making specific to downstream ML systems. Because of the gap between policy-making and execution, administrators usually give little consideration to the practical feasibility of policies whose implementation it is known AI will be employed. I expect new approaches will require different team compositions and a redistribution of power across actors.

Political economy of AI production and consumption: Dominant modes of 'doing AI' require vast amounts of computing power and data storage, which are only available to a handful of oligopolistic corporations that enjoy outsize economic and political power [217, 316]. Furthermore, they depend on tremendous amounts of labor, often done in poor conditions at the peripheries, outside the view of citizens in the wealthy core regions [e.g., 274]. Marginalized communities bear the brunt of social and environmental costs while dominant ones reap the benefits [24]. Researchers would do well to develop alternative modes of doing AI that are less resource-intensive and less exploitative of labor. Such forms of AI would allow for more democratic control and oversight, both at the point of production *and* consumption by workers and citizens alike.

6.6 Concluding Remarks: Politicizing Design, or Design and Real Politics

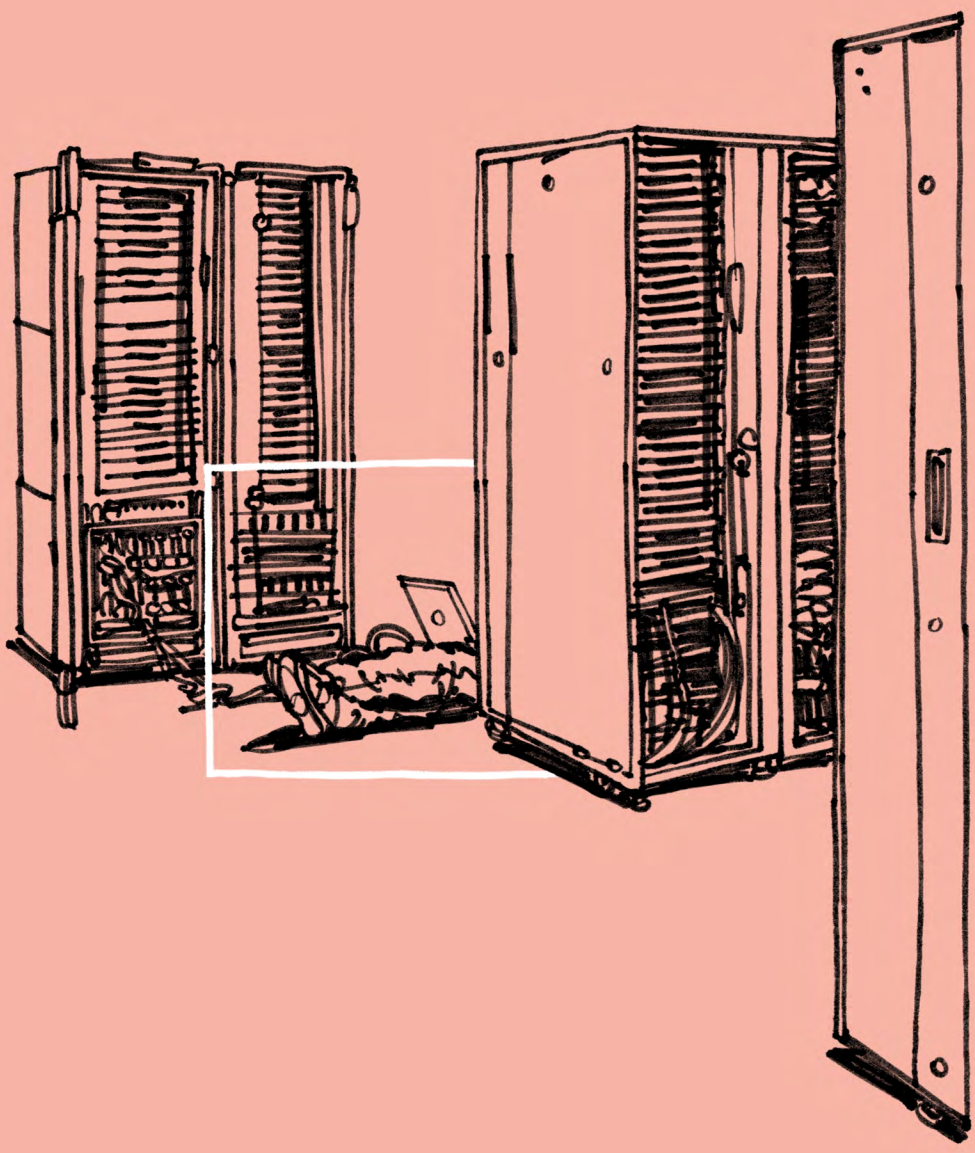
Design ethics has received considerable attention in recent years—spurred on at least in part by liberal panic over the twin shocks of the Brexit referendum vote (June 2016) and the Trump election (November 2016) and the apparent role weaponized social media played in both. See, for example, the Facebook–Cambridge Analytica data scandal [54]. This concern has almost certainly been further fueled by the incessant mutually reinforcing cycles of hype and "critihype" about AI [351].

Despite the increased focus on the social impact of information technology on society and a look towards ethics as a means by which better outcomes can be guaranteed [e.g., 36], the question of power structures, of who does what to whom for whose benefit (in short: of *politics*) tends to receive less explicit attention. This ignorance of politics is even the case for much of participatory design, which ostensibly has a liberatory and democratic agenda but, more often than not, is easily coopted [52]. To be clear, I do not consider myself immune to these shortcomings.

Given the typical background of designers in the Global North's developed cores—including this author's own—it should not surprise us that political leanings tend to be on the liberal and progressive sides. With this background come particular mental models of how democratic politics functions or should function. These models tend to presume that disagreements stem from conflicting values. Furthermore, they idealize rational deliberation as a suitable response. For all the talk of reflexivity in design, the forms of democracy that design outcomes support or undermine are infrequently questioned.

There is, in other words, work to be done on educating ourselves about political philosophy as design researchers. Equally, contributions could seek to help design practitioners orient themselves to the various models of democracy that exist. For example, as Chapter 5 shows, metaphors constructed from political theory can help to push design practice out of its nonideological comfort zone.

I have my political preferences, as has become evident throughout this thesis. Nevertheless, the point is, first and foremost, for designers to become aware of themselves as players who occupy positions on a political field. To engage with what Geuss has termed "real politics" [128]. Lack of such awareness makes designers too easy targets for manipulation and coercion by hegemonic actors. We should not fall into the trap of what Fisher calls "capitalist realism" [107]—uncritically accepting current circumstances as without alternative, the dominant view of postpolitical neoliberal technocracy of the recent past [155]. The task is to broaden our political horizons instead and, in so doing, to politicize design practice itself.



Bibliography

[1] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC Canada: ACM, Apr. 2018, pp. 1–18. DOI: 10/gfzzgc.

[2] Adadi, A. and Berrada, M. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10/gfvb5g.

[3] Aizenberg, E. and Hoven, J. van den. "Designing for human rights in AI." In: *Big Data & Society* 7.2 (July 2020), p. 2053951720949566. DOI: 10/gg8q49.

[4] Aler Tubella, A., Theodorou, A., Dignum, V., and Michael, L. "Contestable Black Boxes." In: *Rules and Reasoning*. Ed. by V. Gutiérrez-Basulto, T. Kliegr, A. Soylu, M. Giese, and D. Roman. Vol. 12173. Cham: Springer International Publishing, 2020, pp. 159–167. URL: https://link.springer.com/10.1007/978-3-030-57977-7_12 (visited on 07/22/2021).

[5] Alfrink, K., Keller, I., Doorn, N., and Kortuem, G. "Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, pp. 1–16. DOI: 10/gr5wcx.

[6] Alfrink, K., Keller, I., Doorn, N., and Kortuem, G. "Tensions in Transparent Urban AI: Designing a Smart Electric Vehicle Charge Point." In: *AI & Society* 38.3 (Mar. 2022), pp. 1049–1065. DOI: 10/gpszwh.

[7] Alfrink, K., Keller, I., Kortuem, G., and Doorn, N. "Contestable AI by Design: Towards a Framework." In: *Minds and Machines* 33.4 (Aug. 2022), pp. 613–639. DOI: 10/gqnjcs.

[8] Alkhatib, A. and Bernstein, M. “Street-level algorithms: A theory at the gaps between policy and decisions.” In: *Conference on Human Factors in Computing Systems - Proceedings*. 2019. DOI: 10/gf9h69.

[9] Almada, M. “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*. 2019, pp. 2–11. DOI: 10/gghft8.

[10] Alvarado, O. and Waern, A. “Towards algorithmic experience: Initial efforts for social media contexts.” In: *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 2018-April. 2018, pp. 1–9. DOI: 10/gf5998.

[11] Ananny, M. and Crawford, K. “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.” In: *New Media and Society* 20.3 (2018), pp. 973–989. DOI: 10/gddxrg.

[12] Anik, A. I. and Bunt, A. “Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency.” In: *CHI*. May 2021. DOI: 10/gs25kh.

[13] Antonelli, P. *States of Design 03: Thinkering*. July 2011. URL: https://www.domusweb.it/en/design/2011/07/04/states-of-design-03-thinkering.html (visited on 03/17/2024).

[14] Applebee, A. N. and Langer, J. A. “Instructional Scaffolding: Reading and Writing as Natural Language Activities.” In: *Language Arts* 60.2 (1983), pp. 168–175. URL: http://www.jstor.org/stable/41961447 (visited on 08/06/2021).

[15] Arnstein, S. R. “A ladder of citizen participation.” In: *Journal of the American Institute of planners* 35.4 (1969), pp. 216–224. DOI: 10/cvct7d.

[16] Auger, J. “Speculative design: crafting the speculation.” In: *Digital Creativity* 24.1 (Mar. 2013), pp. 11–35. DOI: 10/gd4q58.

[17] Bardzell, J. and Bardzell, S. “What is "critical" about critical design?” In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris France: ACM, Apr. 2013, pp. 3297–3306. DOI: 10/ggc5s7.

[18] Bardzell, J., Bardzell, S., and Stolterman, E. “Reading critical designs: supporting reasoned interpretations of critical design.” In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Toronto Ontario Canada: ACM, Apr. 2014, pp. 1951–1960. DOI: 10/f3nnk2.

[19] Bardzell, S., Bardzell, J., Forlizzi, J., Zimmerman, J., and Antanitis, J. "Critical design and critical theory: the challenge of designing for provocation." In: *Proceedings of the Designing Interactive Systems Conference on - DIS '12*. Newcastle Upon Tyne, United Kingdom: ACM Press, 2012, p. 288. DOI: 10/ggpv92.

[20] Barredo Arrieta, A. et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." In: *Information Fusion* 58 (June 2020), pp. 82–115. DOI: 10/ggqs5w.

[21] Baumer, E. P. S., Blythe, M., and Tanenbaum, T. J. "Evaluating Design Fiction: The Right Tool for the Job." In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. Eindhoven Netherlands: ACM, July 2020, pp. 1901–1913. DOI: 10/ghnnv6.

[22] Bayamlıoğlu, E. "The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"." In: *Regulation and Governance* (2021). DOI: 10/gj7sk9.

[23] Beck, J. and Ekbia, H. "The theory-practice gap as generative metaphor." In: *Conf Hum Fact Comput Syst Proc*. Vol. 2018-April. Association for Computing Machinery, 2018. DOI: 10/grz7gb.

[24] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. DOI: 10/gh677h.

[25] Benjamin, J. J., Berger, A., Merrill, N., and Pierce, J. "Machine Learning Uncertainty as a Design Material: A Post-Phenomenological Inquiry." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, May 2021, pp. 1–14. DOI: 10/gksmbj.

[26] Benjamin, J. J. et al. "The Entoptic Field Camera as Metaphor-Driven Research-through-Design with AI Technologies." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–19. DOI: 10/gsk4hj.

[27] Bhatt, U. et al. "Explainable machine learning in deployment." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Jan. 2020. DOI: 10/ghppt9.

[28] Binns, R. "Algorithmic Accountability and Public Reason." In: *Philosophy and Technology* 31.4 (2018), pp. 543–556. DOI: 10/gd89cv.

[29] Binns, R. and Gallo, V. *An overview of the Auditing Framework for Artificial Intelligence and its core components*. Mar. 2019. URL: https://ico.org.uk/about-the-ico/media-centre/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/ (visited on 08/15/2022).

[30] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and 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*. Montreal QC Canada: ACM, Apr. 2018, pp. 1–14. DOI: 10/cvcp.

[31] Bleecker, J. *Design Fiction: A Short Essay on Design, Science, Fact and Fiction*. Mar. 2009. URL: https://blog.nearfuturelaboratory.com/2009/03/17/design-fiction-a-short-essay-on-design-science-fact-and-fiction/ (visited on 08/05/2022).

[32] Bloodworth, A. "Using camera cars to assess the engineering impact of tsunamis on buildings." In: *Proceedings of the Institution of Civil Engineers - Civil Engineering* 168.4 (Nov. 2015), pp. 150–150. DOI: 10/gqmddc.

[33] Blythe, M. "Research through design fiction: narrative in real and imaginary abstracts." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. New York, NY, USA: Association for Computing Machinery, Apr. 2014, pp. 703–712. DOI: 10/gfkwbs.

[34] Bovens, M. and Zouridis, S. "From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control." In: *Public Administration Review* 62.2 (2002), pp. 174–184. DOI: 10/d77d8s.

[35] Bowers, J. "The logic of annotated portfolios: communicating the value of 'research through design'." In: *Proceedings of the Designing Interactive Systems Conference*. Newcastle Upon Tyne United Kingdom: ACM, June 2012, pp. 68–77. DOI: 10/ggzk3v.

[36] Bowles, C. *Future ethics*. East Sussex, United Kingdom: NowNext Press, 2018.

[37] Bratteteig, T. and Verne, G. "Does AI make PD obsolete?: exploring challenges from artificial intelligence to participatory design." In: *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2*. Hasselt and Genk Belgium: ACM, Aug. 2018, pp. 1–5. DOI: 10/ghsn84.

[38] Braun, M., Bleher, H., and Hummel, P. "A Leap of Faith: Is There a Formula for "Trustworthy" AI?" In: *Hastings Center Report* 51.3 (May 2021), pp. 17–22. DOI: 10.1002/hast.1207.

[39] Braun, V. and Clarke, V. "Can I use TA? Should I use TA? Should I *not* use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches." In: *Counselling and Psychotherapy Research* 21.1 (Mar. 2021), pp. 37–47. DOI: 10/ghf388.

[40] Braun, V. and Clarke, V. "Conceptual and design thinking for thematic analysis." In: *Qualitative Psychology* 9.1 (Feb. 2022), pp. 3–26. DOI: 10/gj2m7c.

[41] Braun, V. and Clarke, V. "Reflecting on reflexive thematic analysis." In: *Qualitative Research in Sport, Exercise and Health* 11.4 (Aug. 2019), pp. 589–597. DOI: 10/gf89jz.

[42] Braun, V. and Clarke, V. *Successful qualitative research: a practical guide for beginners*. Los Angeles: SAGE, 2013.

[43] Braun, V. and Clarke, V. *Thematic analysis: a practical guide to understanding and doing*. 1st ed. Thousand Oaks: SAGE Publications, 2021.

[44] Braun, V. and Clarke, V. "Thematic analysis." In: *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* Ed. by H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher. Washington: American Psychological Association, 2012, pp. 57–71. DOI: 10.1037/13620-004.

[45] Braun, V. and Clarke, V. "Using thematic analysis in psychology." In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10/fswdcx.

[46] Brauneis, R. and Goodman, E. P. "Algorithmic transparency for the smart city." In: *Yale JL & Tech.* 20 (2018), p. 103.

[47] Brkan, M. "Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond." In: *International Journal of Law and Information Technology* 27.2 (June 2019), pp. 91–121. DOI: 10/gf33xn.

[48] Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., and Vaithianathan, R. "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, May 2019, pp. 1–12. DOI: 10/gjgz67.

[49] Bruno, R. and Nurchis, M. "Efficient data collection in multimedia vehicular sensing platforms." In: *Pervasive and Mobile Computing* 16 (Jan. 2015), pp. 78–95. DOI: 10/f6wx2c.

[50] Bunt, A., Lount, M., and Lauzon, C. "Are explanations always important?: a study of deployed, low-cost intelligent interactive systems." In: *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces - IUI '12*. Lisbon, Portugal: ACM Press, 2012, p. 169. DOI: 10/ghv472.

[51] Burrell, J. "How the machine 'thinks': Understanding opacity in machine learning algorithms." In: *Big Data and Society* 3.1 (2016), pp. 1–12. DOI: 10/gcd3mk.

[52] Busch, O. von and Palmås, K. "Design is… corrupting." In: *The Design Journal* 26.3 (May 2023), pp. 376–379. DOI: 10/gr7t6z.

[53] Caduff, C. "Targets in the Cloud: On Transparency and Other Shadows." In: *Science Technology and Human Values* 42.2 (2017), pp. 315–319. DOI: 10/gmv782.

[54] Cadwalladr, C. and Graham-Harrison, E. "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach." In: *The Guardian* (Mar. 2018). URL: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (visited on 11/07/2023).

[55] Capel, T. and Brereton, M. "What is Human-Centered about Human-Centered AI? A Map of the Research Landscape." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, pp. 1–23. DOI: 10/gr6q26.

[56] Cardullo, P. and Kitchin, R. *Being a 'citizen' in the smart city: Up and down the scaffold of smart citizen participation*. May 2017. DOI: 10.31235/osf.io/v24jn.

[57] Cavalcante Siebert, L. et al. "Meaningful human control: actionable properties for AI system development." In: *AI and Ethics* (May 2022). DOI: 10/gp6zkx.

[58] Centivany, A. and Glushko, B. “"Popcorn tastes good": Participatory policymaking and Reddit’s "AMAgeddon"." In: *Conference on Human Factors in Computing Systems - Proceedings* (2016), pp. 1126–1137. DOI: 10/ghcvmt.

[59] Chalmers, M. and Galani, A. “Seamful interweaving: heterogeneity in the theory and design of interactive systems.” In: *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*. DIS ’04. New York, NY, USA: Association for Computing Machinery, Aug. 2004, pp. 243–252. DOI: 10/bwscp3.

[60] Cheng, H.-F. et al. “Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders.” In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, Paper 559. URL: https://doi.org/10.1145/3290605.3300789.

[61] Chiusi, F., Fischer, S., Kayser-Bril, N., and Spielkamp, N. *Automating Society Report 2020*. Tech. rep. Algorithm Watch, Oct. 2020. URL: https://automatingsociety.algorithmwatch.org.

[62] Christman, J. *The Politics of Persons: Individual Autonomy and Socio-historical Selves*. Cambridge: Cambridge University Press, 2009. DOI: 10.1017/CBO9780511635571.

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

[64] Collingridge, D. *The social control of technology*. London New York: Frances Pinter St. Martin’s press, 1982.

[65] Cowgill, B. and Tucker, C. “Algorithmic Bias: A Counterfactual Perspective.” In: *Working Paper: NSFTrustworthy Algorithms.* (2017), p. 3. URL: http://trustworthy-algorithms.org/whitepapers/Bo%20Cowgill.pdf.

[66] Crawford, K. “Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics.” In: *Science, Technology, & Human Values* 41.1 (Jan. 2016), pp. 77–92. DOI: 10/gddv8j.

[67] Crawford, K. et al. *AI Now 2019 Report*. Tech. rep. New York: AI Now Institute, 2019. URL: https://ainowinstitute.org/AI_Now_2019_Report.html.

[68] Cugurullo, F. “Urban Artificial Intelligence: From Automation to Autonomy in the Smart City.” In: *Frontiers in Sustainable Cities* 2 (July 2020), pp. 1–14. URL: https://www.frontiersin.org/articles/10.3389/frsc.2020.00038 (visited on 09/20/2023).

[69] Davis, J. “Design methods for ethical persuasive computing.” In: *Proceedings of the 4th international conference on persuasive technology*. Persuasive ’09. New York, NY, USA: Association for Computing Machinery, 2009. DOI: 10/ft9p4f.

[70] De Laat, P. B. “Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?” In: *Philosophy and Technology* 31.4 (2018), pp. 525–541. DOI: 10/gf5997.

[71] De Laat, P. B. “The disciplinary power of predictive algorithms: a Foucauldian perspective.” In: *Ethics and Information Technology* 21.4 (2019), pp. 319–329. DOI: 10/ggjd4g.

[72] Desai, S. and Twidale, M. “Metaphors in Voice User Interfaces: A Slippery Fish.” In: *ACM Transactions on Computer-Human Interaction* 0.0 (July 2023), pp. 1–40. DOI: 10/gsnxcc.

[73] Descampe, A., Massart, C., Poelman, S., Standaert, F.-X., and Standaert, O. “Automated news recommendation in front of adversarial examples and the technical limits of transparency in algorithmic accountability.” In: *AI & SOCIETY* (Mar. 2021). DOI: 10/gjf7jj.

[74] DiSalvo, C. “Design and the Construction of Publics.” In: *Design issues* 25.1 (2009), pp. 48–63. DOI: 10/bjgcd6.

[75] DiSalvo, C. “Design, Democracy and Agonistic Pluralism.” In: *DRS Biennial Conference Series*. Montreal, Canada: Design Research Society, July 2010. URL: https://edu.nl/vwm8x.

[76] Dobbe, R., Krendl Gilbert, T., and Mintz, Y. “Hard choices in artificial intelligence.” In: *Artificial Intelligence* 300 (Nov. 2021), p. 103555. DOI: 10/gmg8pd.

[77] Döbelt, S., Jung, M., Busch, M., and Tscheligi, M. “Consumers’ privacy concerns and implications for a privacy preserving Smart Grid architecture—Results of an Austrian study.” In: *Energy Research & Social Science* 9 (Sept. 2015), pp. 137–145. DOI: 10/gmxmzk.

[78] Dolin, C. et al. “Unpacking perceptions of data-driven inferences underlying online targeting and personalization.” In: *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), pp. 1–12. DOI: 10/gf3cvq.

[79] Dorst, K. *Frame Innovation: Create New Thinking by Design*. Illustrated edition. Cambridge, Massachusetts: The MIT Press, Mar. 2015.

[80] Dorst, K. and Cross, N. “Creativity in the design process: co-evolution of problem–solution.” In: *Design Studies* 22.5 (Sept. 2001), pp. 425–437. DOI: 10/fgw87x.

[81] Dourish, P. “Algorithms and their others: Algorithmic culture in context.” In: *Big Data and Society* 3.2 (2016), pp. 1–11. DOI: 10/gcdx9q.

[82] Dourish, P. “HCI and environmental sustainability: the politics of design and the design of politics.” In: *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. DIS ’10. New York, NY, USA: Association for Computing Machinery, Aug. 2010, pp. 1–10. DOI: 10/b88vcs.

[83] Dourish, P. “What we talk about when we talk about context.” In: *Personal and Ubiquitous Computing* 8.1 (Feb. 2004), pp. 19–30. DOI: 10/cfdfd6.

[84] Dove, G. and Fayard, A.-L. “Monsters, Metaphors, and Machine Learning.” In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–17. DOI: 10/grdb57.

[85] Dove, G., Halskov, K., Forlizzi, J., and Zimmerman, J. “UX Design Innovation: Challenges for Working with Machine Learning as a Design Material.” In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI ’17* 2017-May (2017), pp. 278–288. DOI: 10/cvvd.

[86] Drobotowicz, K., Kauppinen, M., and Kujala, S. “Trustworthy AI Services in the Public Sector: What Are Citizens Saying About It?” In: *Requirements Engineering: Foundation for Software Quality*. Ed. by F. Dalpiaz and P. Spoletini. Vol. 12685. Cham: Springer International Publishing, 2021, pp. 99–115. DOI: 10.1007/978-3-030-73128-1_7.

[87] Dubberly, H. *Why we should stop describing design as “problem solving”*. Oct. 2022. URL: https://edu.nl/nt7b8 (visited on 11/07/2022).

[88] Dunleavy, P., Margetts, H., Bastow, S., and Tinkler, J. “New Public Management Is Dead—Long Live Digital-Era Governance.” In: *Journal of Public Administration Research and Theory* 16.3 (July 2006), pp. 467–494. DOI: 10/cc262q.

[89] Dunn, P. T. “Participatory Infrastructures: The Politics of Mobility Platforms.” In: *Urban Planning* 5.4 (Dec. 2020), pp. 335–346. DOI: 10/ghqjh4.

[90] Dunne, A. and Raby, F. *Speculative everything: design, fiction, and social dreaming*. Cambridge, Massachusetts ; London: The MIT Press, 2013.

[91] Edwards, L. and Veale, M. “Enslaving the Algorithm: From a “Right to an Explanation” to a “Right to Better Decisions”?” In: *IEEE Security & Privacy* 16.3 (May 2018), pp. 46–54. DOI: 10/gdz29v.

[92] Ehn, P. “Participation in design things.” In: *Proceedings of the tenth anniversary conference on participatory design 2008*. PDC ’08. USA: Indiana University, 2008, pp. 92–101.

[93] Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. “Expanding Explainability: Towards Social Transparency in AI systems.” In: May 2021. DOI: 10/gksktj.

[94] Eiband, M., Buschek, D., Kremer, A., and Hussmann, H. “The impact of placebic explanations on trust in intelligent systems.” In: *Conference on Human Factors in Computing Systems - Proceedings* (2019). DOI: 10/gmv78z.

[95] Elkin-Koren, N. “Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence.” In: *Big Data & Society* 7.2 (July 2020), p. 205395172093229. DOI: 10/gg8v9r.

[96] Eslami, M., Vaccaro, K., Lee, M. K., On, A. E. B., Gilbert, E., and Karahalios, K. “User attitudes towards algorithmic opacity and transparency in online reviewing platforms.” In: *Conference on Human Factors in Computing Systems - Proceedings*. 2019, pp. 1–14. DOI: 10/gmv78r.

[97] Eslami, M. et al. “Communicating algorithmic process in online behavioral advertising.” In: *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), pp. 1–13. DOI: 10/cxrf.

[98] Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin’s Press, Jan. 2018.

[99] Fabianek, P., Will, C., Wolff, S., and Madlener, R. “Green and regional? A multi-criteria assessment framework for the provision of green electricity for electric vehicles in Germany.” In: *Transportation Research Part D: Transport and Environment* 87 (Oct. 2020), p. 102504. DOI: 10/gmxmz4.

[100] Fabrocini, F. and Terzidis, K. “Re-framing AI: An AI Product Designer Perspective.” In: *Techné: Research in Philosophy and Technology* 25.3 (2021), pp. 407–433. DOI: 10/gnq25r.

[101] Fan, G., Zhao, Y., Guo, Z., Jin, H., Gan, X., and Wang, X. “Towards Fine-Grained Spatio-Temporal Coverage for Vehicular Urban Sensing Systems.” In: *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. Vancouver, BC, Canada: IEEE, May 2021, pp. 1–10. DOI: 10/gmr355.

[102] Fatima, S., Desouza, K. C., Buck, C., and Fielt, E. “Public AI canvas for AI-enabled public value: A design science approach.” In: *Government Information Quarterly* 39.4 (Oct. 2022), p. 101722. DOI: 10/gqmc79.

[103] Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. “Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns.” In: *Big Data & Society* 6.1 (Jan. 2019), p. 205395171986054. DOI: 10/gf483k.

[104] Ferri, G., Bardzell, J., Bardzell, S., and Louraine, S. “Analyzing critical designs: categories, distinctions, and canons of exemplars.” In: *Proceedings of the 2014 conference on Designing interactive systems*. Vancouver BC Canada: ACM, June 2014, pp. 355–364. DOI: 10/gnkpkh.

[105] Filonik, J. “We Are the Champions: the Role of Agonistic Metaphor in the Political Discourse of Classical Greece.” In: *The agon in classical literature: studies in honour of Professor Chris Carey*. Ed. by M. Edwards, A. Eustathiu, I. Karamanu, E. Bolonakē, and C. Carey. Bulletin of the Institute of Classical Studies 145. London: Institute of Classical Studies, School of Advanced Study, University of London Press, 2022. DOI: 10.14296/wkue3508.

[106] Fine Licht, K. de and Fine Licht, J. de. “Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy.” In: *AI & SOCIETY* 35.4 (Dec. 2020), pp. 917–926. DOI: 10/ghh5p3.

[107] Fisher, M. *Capitalist realism: is there no alternative?* Zero books. Winchester, UK Washington, USA: Zero Books, 2009.

[108] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. SSRN Scholarly Paper. Rochester, NY, Jan. 2020. DOI: 10.2139/ssrn.3518482.

[109] Flügge, A. A. “Perspectives from Practice: Algorithmic Decision-Making in Public Employment Services.” In: *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. Virtual Event USA: ACM, Oct. 2021, pp. 253–255. DOI: 10/gnhcvm.

[110] Forlano, L. and Mathew, A. “From Design Fiction to Design Friction: Speculative and Participatory Design of Values-Embedded Urban Technology.” In: *Journal of Urban Technology* 21.4 (2014), pp. 7–24. DOI: 10/gf65fb.

[111] Foth, M. “Participatory urban informatics: towards citizen- *ability*.” In: *Smart and Sustainable Built Environment* 7.1 (Apr. 2018), pp. 4–19. DOI: 10/gf4mpm.

[112] M. Foth, M. Brynskov, and T. Ojala, eds. *Citizen’s Right to the Digital City*. Singapore: Springer Singapore, 2015. DOI: 10.1007/978-981-287-919-6.

[113] Foth, M., Mann, M., Bedford, L., Fieuw, W., and Walters, R. “A capitalocentric review of technology for sustainable development: The Case for more-than-human design.” In: *Global Information Society Watch 2020: Technology, the environment and a sustainable world: Responses from the global South* (2021), pp. 78–82.

[114] Franssen, M. “Design for Values and Operator Roles in Sociotechnical Systems.” In: *Handbook of Ethics, Values, and Technological Design*. Ed. by J. van den Hoven, P. E. Vermaas, and I. van de Poel. Dordrecht: Springer Netherlands, 2015, pp. 117–149. DOI: 10.1007/978-94-007-6970-0_8.

[115] Frauenberger, C. “Critical Realist HCI.” In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. San Jose California USA: ACM, May 2016, pp. 341–351. DOI: 10/chg6.

[116] Frauenberger, C., Foth, M., and Fitzpatrick, G. “On scale, dialectics, and affect: pathways for proliferating participatory design.” In: *Proceedings of the 15th Participatory Design Conference: Full Papers - Volume 1*. Hasselt and Genk Belgium: ACM, Aug. 2018, pp. 1–13. DOI: 10/gmzc2n.

[117] Frendo, O., Gaertner, N., and Stuckenschmidt, H. “Real-time smart charging based on precomputed schedules.” In: *IEEE Transactions on Smart Grid* 10.6 (2019), pp. 6921–6932. DOI: 10/ghcjvr.

[118] Fryer, T. “A critical realist approach to thematic analysis: producing causal explanations.” In: *Journal of Critical Realism* 21.4 (Aug. 2022), pp. 365–384. DOI: 10/grwcs3.

[119] Galloway, A. and Caudwell, C. “Speculative design as research method: From answers to questions and “staying with the trouble”.” In: *Undesign: Critical Practices at the Intersection of Art and Design*. Taylor and Francis, 2018, pp. 85–96.

[120] Gaver, B. and Bowers, J. “Annotated portfolios.” In: *Interactions* 19.4 (July 2012), pp. 40–49. DOI: 10/gft9qv.

[121] Gebru, T. et al. *Datasheets for Datasets*. Dec. 2021. DOI: 10.48550/arXiv.1803.09010.

[122] Geels, F. W. “A socio-technical analysis of low-carbon transitions: introducing the multi-level perspective into transport studies.” In: *Journal of Transport Geography* 24 (2012), pp. 471–482. DOI: 10/f4b32f.

[123] Gemeente Amsterdam. *Een Digitale Stad voor én van iedereen*. Mar. 2019. URL: https://assets.amsterdam.nl/publish/pages/964754/agenda_digitale_stad_v3.pdf (visited on 09/24/2021).

[124] Gemeente Amsterdam. *Laad-update Amsterdam Q2 2021*. Tech. rep. Gemeente Amsterdam, Aug. 2021. URL: https://edu.nl/k7bv6 (visited on 09/30/2021).

[125] Gemeente Amsterdam. *Nota van Beantwoording Milieuzone Amsterdam 2020*. Tech. rep. Gemeente Amsterdam, Apr. 2020. URL: https://assets.amsterdam.nl/publish/pages/850790/nota_van_beantwoording_milieuzone_amsterdam_2020.pdf.

[126] Genus, A. and Stirling, A. “Collingridge and the dilemma of control: Towards responsible and accountable innovation.” In: *Research Policy* 47.1 (Feb. 2018), pp. 61–69. DOI: 10/gcs7sn.

[127] Geuens, J., Geurts, L., Swinnen, T. W., Westhovens, R., Van Mechelen, M., and Abeele, V. V. “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*. Hasselt and Genk Belgium: ACM, Aug. 2018, pp. 1–5. DOI: 10/gmfhf5.

[128] Geuss, R. *Philosophy and real politics*. Princeton: Princeton University Press, 2008.

[129] Giaccardi, E. “Histories and futures of research through design: From prototypes to connected things.” In: *International Journal of Design* 13.3 (2019), pp. 139–155. URL: http://www.ijdesign.org/index.php/IJDesign/article/view/3192/875 (visited on 10/10/2023).

[130] Gilbert, T. K., Lambert, N., Dean, S., Zick, T., Snoswell, A., and Mehta, S. “Reward Reports for Reinforcement Learning.” In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 84–130. DOI: 10/gs9cnh.

[131] Goodman, B. “Economic Models of (Algorithmic) Discrimination.” In: June 2016.

[132] Gorski, P. S. ““What is Critical Realism? And Why Should You Care?”” In: *Contemporary Sociology: A Journal of Reviews* 42.5 (Sept. 2013), pp. 658–670. DOI: 10/gfvrfz.

[133] Grandinetti, J. “Examining embedded apparatuses of AI in Facebook and TikTok.” In: *AI & SOCIETY* (Sept. 2021). DOI: 10/gmzcb8.

[134] Gray, C. M. “"It’s More of a Mindset Than a Method": UX Practitioners’ Conception of Design Methods.” In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. New York, NY, USA: Association for Computing Machinery, May 2016, pp. 4044–4055. DOI: 10/ghbvk9.

[135] Gray, C. M. and Kou, Y. “UX Practitioners’ Engagement with Intermediate-Level Knowledge.” In: *Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems*. DIS ’17 Companion. New York, NY, USA: Association for Computing Machinery, June 2017, pp. 13–17. DOI: 10/grfr9p.

[136] Green, B. and Viljoen, S. “Algorithmic realism: expanding the boundaries of algorithmic thought.” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona Spain: ACM, Jan. 2020, pp. 19–31. DOI: 10/ggjpcj.

[137] Green, S. D., Kao, C.-C., and Larsen, G. D. “Contextualist Research: Iterating between Methods While Following an Empirically Grounded Approach.” In: *Journal of Construction Engineering and Management* 136.1 (Jan. 2010), pp. 117–126. DOI: 10/cx5fsm.

[138] Grimmelikhuijsen, S. and Meijer, A. “Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response.” In: *Perspectives on Public Management and Governance* 5.3 (Sept. 2022), pp. 232–242. DOI: 10/gr49r2.

[139] Günther, M. and Kasirzadeh, A. “Algorithmic and human decision making: for a double standard of transparency.” In: *AI & SOCIETY* (Apr. 2021). DOI: 10/gkbk5b.

[140] Hamilton, A. “Metaphor in theory and practice: the influence of metaphors on expectations.” In: *ACM Journal of Computer Documentation* 24.4 (Nov. 2000), pp. 237–253. DOI: 10/ckdrsn.

[141] Haraway, D. J. *Staying with the trouble: making kin in the Chthulucene*. Experimental futures: technological lives, scientific arts, anthropological voices. Durham: Duke University Press, 2016.

[142] Harvey, D. “The right to the city.” In: *International Journal of Urban and Regional Research* 27.4 (Dec. 2003), pp. 939–941. DOI: 10/cmrn37.

[143] Hekkert, P. and Cila, N. “Handle with care! Why and how designers make use of product metaphors.” In: *Design Studies* 40 (Sept. 2015), pp. 196–217. DOI: 10/gc8zwb.

[144] Helwig, J. *‘Ik wil dat iemand zegt dat ik geen fraudeur ben’*. June 2023. URL: https://edu.nl/w94pq (visited on 11/09/2023).

[145] Henin, C. and Le Métayer, D. “A framework to contest and justify algorithmic decisions.” In: *AI and Ethics* 1.4 (Nov. 2021), pp. 463–476. DOI: 10/gmbqw6.

[146] Henin, C. and Le Métayer, D. “Beyond explainability: justifiability and contestability of algorithmic decision systems.” In: *AI & SOCIETY* 37.4 (Dec. 2022), pp. 1397–1410. DOI: 10/gmg8pf.

[147] Henwood, K. and Pidgeon, N. “Beyond the qualitative paradigm: A framework for introducing diversity within qualitative psychology.” In: *Journal of Community & Applied Social Psychology* 4.4 (Oct. 1994), pp. 225–238. DOI: 10/c94p2d.

[148] Hildebrandt, M. “Law as Information in the Era of Data-Driven Agency.” In: *Modern Law Review* 79.1 (2016), pp. 1–30. DOI: 10/f8dkk9.

[149] Hildebrandt, M. “Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning.” In: *Theoretical Inquiries in Law* 20.1 (Mar. 2019), pp. 83–121. DOI: 10/gfz335.

[150] Hill, D. *Dark matter and trojan horses: a strategic design vocabulary*. First edition. Moscow: Strelka Press, 2012.

[151] Hillgren, P.-A., Seravalli, A., and Emilson, A. “Prototyping and infrastructuring in design for social innovation.” In: *CoDesign* 7.3-4 (2011), pp. 169–183. DOI: 10/f234v3.

[152] Himmelreich, J. “Against “Democratizing AI”.” In: *AI & SOCIETY* (Jan. 2022). DOI: 10/gr95d5.

[153] Hirsch, T., Merced, K., Narayanan, S., Imel, Z. E. Z. E., and Atkins, D. C. D. C. "Designing contestability: Interaction design, machine learning, and mental health." In: *DIS 2017 - Proceedings of the 2017 ACM Conference on Designing Interactive Systems*. ACM Press, 2017, pp. 95–99. DOI: 10/gddxqb.

[154] Hirschman, A. O. *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Vol. 25. Harvard university press, 1970.

[155] Hochuli, A., Hoare, G., and Cunliffe, P. *The End of the End of History: Politics in the Twenty-First Century*. Ridgefield: Zero Books, June 2021.

[156] Hoepman, J.-H. "Transparency Is The Perfect Cover-Up (If The Sun Does Not Shine)." In: *BEING PROFILED*. Ed. by E. Bayamlioglu, I. Baraliuc, L. A. W. Janssens, and M. Hildebrandt. Amsterdam University Press, Dec. 2019, pp. 46–51. DOI: 10.1515/9789048550180-009.

[157] Hollanek, T. "AI transparency: a matter of reconciling design with critique." In: *AI & SOCIETY* (Nov. 2020). DOI: 10/gmzcgp.

[158] Holmquist, L. E. "Intelligence on tap: artificial intelligence as a new design material." In: *Interactions* 24.4 (June 2017), pp. 28–33. DOI: 10/gc8zwk.

[159] C. Hood and D. Heald, eds. *Transparency: the key to better governance?* Proceedings of the British Academy 135. Oxford ; New York: Oxford University Press, 2006.

[160] Höök, K. and Löwgren, J. "Characterizing Interaction Design by Its Ideals: A Discipline in Transition." In: *She Ji: The Journal of Design, Economics, and Innovation* 7.1 (Mar. 2021), pp. 24–40. DOI: 10/gjtbqv.

[161] Höök, K. and Löwgren, J. "Strong Concepts: Intermediate-Level Knowledge in Interaction Design Research." In: *ACM Transactions on Computer-Human Interaction* 19.3 (2012), pp. 1–18. DOI: 10/f225d4.

[162] Höök, K. et al. "A Glass Box Approach to Adaptive Hypermedia." In: *Adaptive Hypertext and Hypermedia*. Ed. by P. Brusilovsky, A. Kobsa, and J. Vassileva. Dordrecht: Springer Netherlands, 1998, pp. 143–170. DOI: 10.1007/978-94-017-0617-9_6.

[163] Howe, B. et al. "Integrative urban AI to expand coverage, access, and equity of urban data." In: *The European Physical Journal Special Topics* 231.9 (July 2022), pp. 1741–1752. DOI: 10/gqmc7g.

[164] Huang, Y. and Kockelman, K. M. “Electric vehicle charging station locations: Elastic demand, station congestion, and network equilibrium.” In: *Transportation Research Part D: Transport and Environment* 78 (Jan. 2020), p. 102179. DOI: 10/gk56dj.

[165] Hutchinson, B. et al. “Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure.” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM, Mar. 2021, pp. 560–575. DOI: 10/gjftws.

[166] Innerarity, D. “Making the black box society transparent.” In: *AI & SOCIETY* 36.3 (Sept. 2021), pp. 975–981. DOI: 10/gkbk5c.

[167] Isbister, K. and Höök, K. “On being supple: in search of rigor without rigidity in meeting new design and evaluation challenges for HCI practitioners.” In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston MA USA: ACM, Apr. 2009, pp. 2233–2242. DOI: 10/frkzx2.

[168] Jackson, S. J., Gillespie, T., and Payette, S. “The Policy Knot: Reintegrating Policy, Practice and Design.” In: *CSCW Studies of Social Computing* (2014), pp. 588–602. DOI: 10/gg5g9w.

[169] Jaeger, M. E. and Rosnow, R. L. “Contextualism and its implications for psychological inquiry.” In: *British Journal of Psychology* 79.1 (Feb. 1988), pp. 63–75. DOI: 10/cx4bvj.

[170] Jasanoff, S. and Kim, S.-H. “Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea.” In: *Minerva* 47.2 (2009), pp. 119–146. DOI: 10/fghmvb.

[171] S. Jasanoff and S.-H. Kim, eds. *Dreamscapes of modernity: sociotechnical imaginaries and the fabrication of power*. Chicago ; London: The University of Chicago Press, 2015.

[172] Jewell, M. “Contesting the decision: living in (and living with) the smart city.” In: *International Review of Law, Computers & Technology* 32.2-3 (Apr. 2018), pp. 210–229. URL: https://www.tandfonline.com/doi/full/10.1080/13600869.2018.1457000 (visited on 09/01/2023).

[173] Jhaver, S., Karpfen, Y., and Antin, J. “Algorithmic anxiety and coping strategies of airbnb hosts.” In: *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 2018-April. 2018, pp. 1–12. DOI: 10/gjnkj9.

[174] Jobin, A., Ienca, M., and Vayena, E. “The global landscape of AI ethics guidelines.” In: *Nature Machine Intelligence* 1.9 (2019), pp. 389–399. DOI: 10/gf73q2.

[175] Johnson, D. W. “Social Interdependence: Interrelationships Among Theory, Research, and Practice.” In: *American Psychologist* 58.11 (Nov. 2003), pp. 934–945. DOI: 10/ftszx2.

[176] Jost, J. T., Federico, C. M., and Napier, J. L. “Political Ideology: Its Structure, Functions, and Elective Affinities.” In: *Annual Review of Psychology* 60.1 (Jan. 2009), pp. 307–337. DOI: 10/dtgrpr.

[177] Kamarinou, D., Millard, C., and Singh, J. “Machine learning with personal data.” In: *Queen Mary School of Law Legal Studies Research Paper* 247 (Nov. 2016), p. 23. URL: https://ssrn.com/abstract=2865811.

[178] Kamols, N., Foth, M., and Guaralda, M. “Beyond engagement theatre: challenging institutional constraints of participatory planning practice.” In: *Australian Planner* 57.1 (Jan. 2021), pp. 23–35. DOI: 10/gmx7g6.

[179] Kariotis, T. and J. Mir, D. “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*. Manizales Colombia: ACM, June 2020, pp. 148–153. DOI: 10/gk48w7.

[180] Katell, M. et al. “Toward situated interventions for algorithmic equity: lessons from the field.” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 45–55. URL: https://doi.org/10.1145/3351095.3372874.

[181] Kemmis, S. and McTaggart, R. “Participatory Action Research: Communicative Action and the Public Sphere.” In: *Educational Action Research* 14.4 (2006), pp. 459–476. DOI: 10/c7h7nm.

[182] Kim, B., Park, J., and Suh, J. “Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information.” In: *Decision Support Systems* 134 (July 2020), p. 113302. DOI: 10/gmxzjn.

[183] Kirby, D. “The Future is Now: Diegetic Prototypes and the Role of Popular Films in Generating Real-world Technological Development.” In: *Social Studies of Science* 40.1 (Feb. 2010), pp. 41–70. DOI: 10/fcn38m.

[184] Kitchin, R. "The ethics of smart cities and urban science." In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2083 (Dec. 2016), p. 20160115. DOI: 10/bs3w.

[185] Kitchin, R. "The real-time city? Big data and smart urbanism." In: *GeoJournal* 79.1 (Feb. 2014), pp. 1–14. DOI: 10/7ds.

[186] Kitchin, R. "Toward a Genuinely Humanizing Smart Urbanism." In: *The Right to the Smart City* (2019), pp. 193–204. DOI: 10/gmv79c.

[187] Kitchin, R., Coletta, C., and Mcardle, G. "Urban informatics, governmentality and the logics of urban control." In: *SocArXiv* (2017), pp. 1–21. DOI: 10/gmv79r.

[188] Kizilcec, R. F. "How much information? Effects of transparency on trust in an algorithmic interface." In: *Conference on Human Factors in Computing Systems - Proceedings*. 2016, pp. 2390–2395. DOI: 10/ggfkc4.

[189] Kluttz, D. N., Kohli, N., and Mulligan, D. K. "Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions." In: *After the digital tornado: Networks, algorithms, humanity*. Ed. by K. Werbach. Cambridge: Cambridge University Press, 2020, pp. 137–152.

[190] Kluttz, D. N. and Mulligan, D. K. "Automated Decision Support Technologies and the Legal Profession." In: *Berkeley Technology Law Journal* 34.3 (2019), p. 853. DOI: 10.15779/Z38154DP7K.

[191] Knutz, E., Markussen, T., and Christensen, P. R. "The Role of Fiction in Experiments within Design, Art & Architecture." In: *Artifact* 3.2 (Dec. 2014), p. 8. DOI: 10/gmrb63.

[192] König, P. D. and Wenzelburger, G. "The legitimacy gap of algorithmic decision-making in the public sector: Why it arises and how to address it." In: *Technology in Society* 67 (Nov. 2021), p. 101688. DOI: 10/gpk2ps.

[193] Koskinen, I., Binder, T., and Redström, J. "Lab, Field, Gallery, and Beyond." In: *Artifact* 2.1 (Apr. 2008), pp. 46–57. DOI: 10/dz3gnf.

[194] Koskinen, I., Zimmerman, J., Binder, T., Redström, J., and Wensveen, S. *Design Research Through Practice*. Boston: Morgan Kaufmann, Jan. 2012. DOI: 10.1016/B978-0-12-385502-2.00013-4.

[195] Kozubaev, S. et al. "Expanding Modes of Reflection in Design Futuring." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, Apr. 2020, pp. 1–15. DOI: 10/gh99ht.

[196] Krafft, P. M. et al. “An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists.” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 772–781. DOI: 10/gh73cg.

[197] Kroes, P., Franssen, M., Poel, I. v. d., and Ottens, M. “Treating socio-technical systems as engineering systems: some conceptual problems.” In: *Systems Research and Behavioral Science* 23.6 (Feb. 2006), pp. 803–814. DOI: 10/dfswsh.

[198] Krogh, P. G. and Koskinen, I. *Drifting by intention: four epistemic traditions from within constructive design research*. Design research foundations. Cham: Springer, 2020.

[199] Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. “Accountable algorithms.” In: *U. Pa. L. Rev.* 165 (2016), p. 633.

[200] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. “Too much, too little, or just right? Ways explanations impact end users’ mental models.” In: *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. San Jose, CA, USA: IEEE, Sept. 2013, pp. 3–10. DOI: 10/gfkmnn.

[201] Lai, Y., Xu, Y., Mai, D., Fan, Y., and Yang, F. “Optimized Large-Scale Road Sensing Through Crowdsourced Vehicles.” In: *IEEE Transactions on Intelligent Transportation Systems* 23.4 (Apr. 2022), pp. 3878–3889. DOI: 10/gqmdc3.

[202] Lakoff, G. *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press, 1987.

[203] Leahu, L. “Ontological Surprises: A Relational Perspective on Machine Learning.” In: *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. Brisbane QLD Australia: ACM, June 2016, pp. 182–186. DOI: 10/gmfbd5.

[204] Lee, M. K., Jain, A., Cha, H. J., Ojha, S., and Kusbit, D. “Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation.” In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–26. DOI: 10.1145/3359284.

[205] Lee, U. and Gerla, M. “A survey of urban vehicular sensing platforms.” In: *Computer Networks* 54.4 (Mar. 2010), pp. 527–544. DOI: 10/bqd52z.

[206] Lefebvre, H. “Le droit à la ville.” In: *L’Homme et la société* 6.1 (1967), pp. 29–35. DOI: 10/ggb9nj.

[207] Leydens, J. A. and Lucena, J. C. *Engineering justice: transforming engineering education and practice*. IEEE PCS Professional Engineering Communication Series. Hoboken, NJ : Piscataway, NJ: John Wiley & Sons ; IEEE Press, 2018.

[208] Lim, Y.-K., Stolterman, E., and Tenenberg, J. “The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas.” In: *ACM Transactions on Computer-Human Interaction* 15.2 (July 2008), pp. 1–27. DOI: 10/d2dsfd.

[209] Lindh, M. “As a Utility – Metaphors of Information Technologies.” In: *Human IT: Journal for Information Technology Studies as a Human Science* 13.2 (May 2016), pp. 47–80. URL: https://humanit.hb.se/article/view/418 (visited on 09/01/2023).

[210] Lipsky, M. *Street-level bureaucracy: dilemmas of the individual in public services*. Russell Sage Foundation, 1983.

[211] Lipsky, M. “Toward a Theory of Street-Level Bureaucracy.” Aug. 1969.

[212] Logler, N., Friedman, B., and Yoo, D. “Metaphor cards: A how-to-guide for making and using a generative metaphorical design toolkit.” In: *DIS - Proc. Des. Interact. Syst. Conf.* Association for Computing Machinery, Inc, 2018, pp. 1373–1386. DOI: 10/gj9m2z.

[213] Lorino, P. “Abduction.” In: *Pragmatism and Organization Studies*. Vol. 1. Oxford University Press, Mar. 2018. DOI: 10.1093/oso/9780198753216.003.0007.

[214] Löwgren, J., Gaver, B., and Bowers, J. “Annotated Portfolios and Other Forms of Intermediate-Level Knowledge.” In: *Interactions* 20.1 (2013), pp. 30–34. DOI: 10/f23xgp.

[215] Löwgren, J. and Stolterman, E. *Thoughtful interaction design: A design perspective on information technology*. Mit Press, 2004.

[216] Lowndes, V. and Paxton, M. “Can agonism be institutionalised? Can institutions be agonised? Prospects for democratic design.” In: *British Journal of Politics and International Relations* 20.3 (2018), pp. 693–710. DOI: 10/gdw7jp.

[217] Luitse, D. and Denkena, W. “The great Transformer: Examining the role of large language models in the political economy of AI.” In: *Big Data & Society* 8.2 (July 2021), p. 205395172110477. DOI: 10/gr2w8k.

[218] Lundberg, S. M. and Lee, S.-I. “A unified approach to interpreting model predictions.” In: *Advances in neural information processing systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[219] Luusua, A. and Ylipulli, J. “Artificial Intelligence and Risk in Design.” In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. Eindhoven Netherlands: ACM, July 2020, pp. 1235–1244. DOI: 10/gg38dj.

[220] Luusua, A. and Ylipulli, J. “Urban AI: Formulating an agenda for the interdisciplinary research of artificial intelligence in cities.” In: *Companion publication of the 2020 ACM designing interactive systems conference*. DIS’ 20 companion. New York, NY, USA: Association for Computing Machinery, 2020, pp. 373–376. DOI: 10/gjr4r6.

[221] Luusua, A., Ylipulli, J., Foth, M., and Aurigi, A. “Urban AI: understanding the emerging role of artificial intelligence in smart cities.” In: *AI & SOCIETY* 38.3 (June 2023), pp. 1039–1044. DOI: 10.1007/s00146-022-01537-5.

[222] Lyons, H., Miller, T., and Velloso, E. “Algorithmic decisions, desire for control, and the preference for human review over algorithmic review.” In: *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. FAccT ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 764–774. DOI: 10/gsb98h.

[223] Lyons, H., Velloso, E., and Miller, T. “Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions.” In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (Apr. 2021), pp. 1–25. DOI: 10/gpnd53.

[224] Madhu, G. M., Vyjayanthi, C., and Modi, C. N. “A Novel Framework for Monitoring Solar PV based Electric Vehicle Community Charging Station and Grid Frequency Regulation using Blockchain.” In: *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. Kanpur, India: IEEE, July 2019, pp. 1–7. DOI: 10/gmxmz3.

[225] Madill, A., Jordan, A., and Shirley, C. “Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies.” In: *British Journal of Psychology* 91.1 (Feb. 2000), pp. 1–20. DOI: 10/bkrsn4.

[226] Mahendran, A. and Vedaldi, A. “Understanding deep image representations by inverting them.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. June 2015.

[227] Malpass, M. “Between Wit and Reason: Defining Associative, Speculative, and Critical Design in Practice.” In: *Design and Culture* 5.3 (Nov. 2013), pp. 333–356. DOI: 10/gc8zsz.

[228] Malpass, M. “Critical Design Practice: Theoretical Perspectives and Methods of Engagement.” In: *The Design Journal* 19.3 (May 2016), pp. 473–489. DOI: 10/gk3frk.

[229] Malpass, M. “Criticism and Function in Critical Design Practice.” In: *Design Issues* 31.2 (Apr. 2015), pp. 59–71. DOI: 10/gc8ztj.

[230] Marda, V. and Narayan, S. “Data in New Delhi’s predictive policing system.” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona Spain: ACM, Jan. 2020, pp. 317–324. DOI: 10/ggjpcs.

[231] Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., and DeTar, C. *Reporting, reviewing, and responding to harassment on twitter*. 2015. DOI: 10.48550/ARXIV.1505.03359.

[232] McCullough, M. *Digital ground: architecture, pervasive computing, and environmental knowing*. 1st paperback ed. Cambridge, Mass: MIT Press, 2005.

[233] McQuillan, D. “People’s Councils for Ethical Machine Learning.” In: *Social Media + Society* 4.2 (Apr. 2018), p. 205630511876830. DOI: 10/gm469q.

[234] Mehta, R., Srinivasan, D., Khambadkone, A. M., Yang, J., and Trivedi, A. “Smart charging strategies for optimal integration of plug-in electric vehicles within existing distribution system infrastructure.” In: *IEEE Transactions on Smart Grid* 9.1 (2018), pp. 299–312. DOI: 10/gcs7rq.

[235] Meilvang, M. L. and Dahler, A. M. “Decision support and algorithmic support: the construction of algorithms and professional discretion in social work.” In: *European Journal of Social Work* 0.0 (Apr. 2022), pp. 1–13. DOI: 10/gqnrfn.

[236] Mendoza, I. and Bygrave, L. A. “The right not to be subject to automated decisions based on profiling.” In: *EU Internet Law*. Springer, 2017, pp. 77–98.

[237] Menéndez-Viso, A. “Black and white transparency: Contradictions of a moral metaphor.” In: *Ethics and Information Technology* 11.2 (2009), pp. 155–162. DOI: 10/bk99rm.

[238] Methnani, L., Aler Tubella, A., Dignum, V., and Theodorou, A. "Let Me Take Over: Variable Autonomy for Meaningful Human Control." In: *Frontiers in Artificial Intelligence* 4 (Sept. 2021). DOI: 10/gnfvvz.

[239] Mhlambi, S. and Tiribelli, S. "Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms." In: *Topoi* 42.3 (July 2023), pp. 867–880. DOI: 10/gr93wp.

[240] Mingardo, G. "Rotterdam, The Netherlands." In: *Parking*. Elsevier, 2020, pp. 133–145. DOI: 10.1016/B978-0-12-815265-2.00008-X.

[241] Mitchell, M. et al. "Model Cards for Model Reporting." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Atlanta GA USA: ACM, Jan. 2019, pp. 220–229. DOI: 10/gftgjg.

[242] Mittelstadt, B. D. B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. "The ethics of algorithms: Mapping the debate." In: *Big Data and Society* 3.2 (2016), p. 205395171667967. DOI: 10/gcdx92.

[243] Moghaddam, Z., Ahmad, I., Habibi, D., and Phung, Q. V. "Smart charging strategy for electric vehicle charging stations." In: *IEEE Transactions on Transportation Electrification* 4.1 (2018), pp. 76–88. DOI: 10/ggwgv7.

[244] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and The PRISMA Group. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." In: *PLoS Medicine* 6.7 (July 2009), e1000097. DOI: 10/bq3jpc.

[245] Mohseni, S. "Toward Design and Evaluation Framework for Interpretable Machine Learning Systems." In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Honolulu HI USA: ACM, Jan. 2019, pp. 553–554. DOI: 10/gqc2b8.

[246] Monno, V. and Khakee, A. "Tokenism or Political Activism? Some Reflections on Participatory Planning." In: *International Planning Studies* 17.1 (Feb. 2012), pp. 85–101. DOI: 10/f233mt.

[247] Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. "From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices." In: *Science and Engineering Ethics* 26 (Dec. 2019), pp. 2141–2168. DOI: 10/ghc8ng.

[248] Morozov, E. *To save everything, click here: the folly of technological solutionism*. First edition. New York: PublicAffairs, 2013.

[249] Mouffe, C. *Agonistics: Thinking The World Politically*. 1st edition. London New York (N.Y.): Verso, July 2013.

[250] Mouffe, C. “Deliberative democracy or agonistic pluralism?” In: *Social Research* 66.3 (1999), p. 745. URL: http://proquest.umi.com/pqdweb?did=46753355&Fmt=7&clientId=58117&RQT=309&VName=PQD.

[251] Mouffe, C. *On the Political*. First Edition. London ; New York: Routledge, June 2005.

[252] Mouffe, C. “Pluralism, dissensus and democratic citizenship.” In: *Education and the good society*. Springer, 2004, pp. 42–53.

[253] Mouffe, C. “Some reflections on an agonistic approach to the public.” In: *Making Things Public: Atmospheres of Democracy*. Ed. by B. Latour and P. Weibel. Cambridge, MA: MIT Press, 2005, pp. 804–807.

[254] Mouffe, C. *The democratic paradox*. London ; New York: Verso, 2000.

[255] Mouffe, C. *The return of the political*. Phronesis. London ; New York: Verso, 1993.

[256] Muldoon, J. “Institutionalizing Radical Democracy: Socialist Republicanism and Democratizing the Economy.” In: *New Political Science* 43.2 (Apr. 2021), pp. 189–207. DOI: 10/grv63q.

[257] Muratori, M. “Integrating EVs in the electricity system.” In: (Nov. 2019). URL: https://www.osti.gov/biblio/1576477.

[258] Murphy, J. W. and Taylor, R. R. “To democratize or not to democratize AI? That is the question.” In: *AI and Ethics* (June 2023). DOI: 10/gtfrg9.

[259] Murray-Rust, D., Nicenboim, I., and Lockton, D. “Metaphors for designers working with AI.” In: *DRS Biennial Conference Series*. Bilbao, Spain, June 2022, pp. 1–20. DOI: https://doi.org/10.21606/drs.2022.667.

[260] Myers West, S. “Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms.” In: *New Media & Society* 20.11 (Nov. 2018), pp. 4366–4383. DOI: 10/gddvzz.

[261] Nicenboim, I., Venkat, S., Rustad, N. L., Vardanyan, D., Giaccardi, E., and Redström, J. “Conversation Starters: How Can We Misunderstand AI Better?” In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA ’23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–4. DOI: 10/gr928h.

[262] Nissenbaum, H. “A contextual approach to privacy online.” In: *Daedalus* 140.4 (Oct. 2011), pp. 32–48. DOI: 10/fjkb7c.

[263] Noordt, C. van and Misuraca, G. “Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union.” In: *Government Information Quarterly* 39.3 (July 2022), p. 101714. DOI: 10/grmx7m.

[264] Norman, D. A. and Stappers, P. J. “DesignX: Complex Sociotechnical Systems.” In: *She Ji: The Journal of Design, Economics, and Innovation* 1.2 (2015), pp. 83–106. DOI: 10/gc3j7r.

[265] Nouws, S., Janssen, M., and Dobbe, R. “Dismantling Digital Cages: Examining Design Practices for Public Algorithmic Systems.” In: *Electronic Government: 21st IFIP WG 8.5 International Conference, EGOV 2022, Linköping, Sweden, September 6–8, 2022, Proceedings*. Berlin, Heidelberg: Springer-Verlag, Sept. 2022, pp. 307–322. DOI: 10/gsqvzf.

[266] Novick, D. G. and Sutton, S. “What is mixed-initiative interaction.” In: *Proceedings of the AAAI spring symposium on computational models for mixed initiative interaction*. Vol. 2. 1997, p. 12.

[267] Obar, J. A. “Sunlight alone is not a disinfectant: Consent and the futility of opening Big Data black boxes (without assistance).” In: *Big Data & Society* 7.1 (Jan. 2020), p. 2053951720935615. DOI: 10/gnt5xk.

[268] Obrenovic, Z. “Design-based research: What we learn when we engage in design of interactive systems.” In: *Interactions* 18.5 (2011), pp. 56–59. DOI: 10/dqs7df.

[269] Ofgem. *Ofgem’s Future Insights Paper 5 - Implications of the transition to electric vehicles*. Tech. rep. Ofgem, July 2018. URL: https://edu.nl/j8rxn (visited on 09/22/2021).

[270] Ortar, N. and Ryghaug, M. “Should all cars be electric by 2025? The electric car debate in Europe.” In: *Sustainability (Switzerland)* 11.7 (2019), pp. 1–16. DOI: 10/gmq5t9.

[271] Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. “Rayyan: A web and mobile app for systematic reviews.” In: *Systematic Reviews* 5.1 (Dec. 2016), p. 210. DOI: 10/gfkdzd.

[272] Ozkaramanli, D., Karahanoğlu, A., and Verbeek, P. “Reflecting on Design Methods and Democratic Technology Development: The Case of Dutch Covid-19 Digital Contact-Tracing Application.” In: *She Ji: The Journal of Design, Economics, and Innovation* 8.2 (2022), pp. 244–269. DOI: 10/gqktmz.

[273] Pau, G. and Tse, R. “Challenges and opportunities in immersive vehicular sensing: Lessons from urban deployments.” In: *Signal Processing: Image Communication* 27.8 (Sept. 2012), pp. 900–908. DOI: 10/f4b22s.

[274] Perrigo, B. *Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer*. Jan. 2023. URL: https://time.com/6247678/openai-chatgpt-kenya-workers/ (visited on 11/13/2023).

[275] Peter, F. *The Grounds of Political Legitimacy*. Oxford, New York: Oxford University Press, Aug. 2023. URL: https://academic.oup.com/book/46053.

[276] Petersen, A. C. M., Christensen, L. R., and Hildebrandt, T. T. “The Role of Discretion in the Age of Automation.” In: *Computer Supported Cooperative Work (CSCW)* 29.3 (June 2020), pp. 303–333. DOI: 10/gjhqh5.

[277] Pine, K. H. and Liboiron, M. “The Politics of Measurement and Action.” In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul Republic of Korea: ACM, Apr. 2015, pp. 3147–3156. DOI: 10/gf65dx.

[278] Ploug, T. and Holm, S. “The four dimensions of contestable AI diagnostics - A patient-centric approach to explainable AI.” In: *Artificial Intelligence in Medicine* 107 (July 2020), p. 101901. DOI: 10/gpk2pk.

[279] Poel, I. van de. “An ethical framework for evaluating experimental technology.” In: *Science and engineering ethics* 22.3 (2016), pp. 667–686. DOI: 10/gdnf49.

[280] Poel, I. van de. “Embedding Values in Artificial Intelligence (AI) Systems.” In: *Minds and Machines* (2020). DOI: 10/ghbm4v.

[281] Popa, E. O., Blok, V., and Wesselink, R. “An Agonistic Approach to Technological Conflict.” In: *Philosophy & Technology* 34.4 (Dec. 2021), pp. 717–737. DOI: 10/gmqjhn.

[282] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. “Manipulating and Measuring Model Interpretability.” In: May 2021. DOI: 10/gksk2v.

[283] Prunkl, C. “Human autonomy in the age of artificial intelligence.” In: *Nature Machine Intelligence* 4.2 (Feb. 2022), pp. 99–101. DOI: 10/gsd2rt.

[284] Rader, E., Cotter, K., and Cho, J. “Explanations as mechanisms for supporting algorithmic transparency.” In: *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 2018-April. 2018, pp. 1–13. DOI: 10/gg28qc.

[285] Raji, I. D. et al. “Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing.” In: *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), pp. 33–44. DOI: 10/ggjpcm.

[286] Ranchordás, S. “Nudging citizens through technology in smart cities.” In: *International Review of Law, Computers & Technology* 34.3 (Sept. 2020), pp. 254–276. DOI: 10/gjjhbh.

[287] Raven, R., Sengers, F., Spaeth, P., Xie, L., Cheshmehzangi, A., and Jong, M. de. “Urban experimentation and institutional arrangements.” In: *European Planning Studies* 27.2 (Feb. 2019), pp. 258–281. DOI: 10.1080/09654313.2017.1393047.

[288] Ribeiro, M. T., Singh, S., and Guestrin, C. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier.” In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. DOI: 10/gfgrbd.

[289] Robertson, S. and Salehi, N. *What If I Don’t Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design*. July 2020. DOI: 10.48550/arXiv.2007.06718.

[290] Rosner, D. K., Kawas, S., Li, W., Tilly, N., and Sung, Y.-C. “Out of Time, Out of Place: Reflections on Design Workshops as a Research Method.” In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. CSCW ’16. New York, NY, USA: Association for Computing Machinery, Feb. 2016, pp. 1131–1141. DOI: 10.1145/2818048.2820021.

[291] Rouvroy, A. “The end(s) of critique : data-behaviourism vs. due-process.” In: *Privacy, Due Process and the Computational Turn. Philosophers of Law Meet Philosophers of Technology*. Ed. by M. Hildebrandt and E. De Vries. Routledge, 2012.

[292] Rubel, A., Castro, C., and Pham, A. K. *Algorithms and autonomy: the ethics of automated decision systems*. Cambridge, United Kingdom ; New York, NY: Cambridge University Press, May 2021.

[293] Rudin, C. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 206–215. DOI: 10/gf4tp9.

[294] Ryan, C. “Emanuela Ceva’s Interactive Justice.” In: *Critical Review of International Social and Political Philosophy* 22.4 (2019), pp. 480–486. DOI: 10/gqcjxt.

[295] Sadowski, J. and Pasquale, F. “The spectrum of control: A social theory of the smart city.” In: *First Monday* 20.7 (2015), pp. 1–22. DOI: 10.5210/fm.v20i7.5903.

[296] Salehi, N., Teevan, J., Iqbal, S., and Kamar, E. “Communicating Context to the Crowd for Complex Writing Tasks.” In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland Oregon USA: ACM, Feb. 2017, pp. 1890–1901. DOI: 10/gmfggq.

[297] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. “Auditing algorithms: Research methods for detecting discrimination on internet platforms.” In: *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. Seattle, WA, USA, May 2014.

[298] Sarra, C. “Put Dialectics into the Machine: Protection against Automatic-decision-making through a Deeper Understanding of *Contestability by Design*.” In: *Global Jurist* 20.3 (Oct. 2020), p. 20200003. DOI: 10/gj7sk6.

[299] Sawhney, N. “Contestations in urban mobility: rights, risks, and responsibilities for Urban AI.” In: *AI & SOCIETY* 38.3 (June 2023), pp. 1083–1098. DOI: 10/gqmc7n.

[300] Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., and Guha, S. “A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare.” In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), pp. 1–41. DOI: 10/gnhcrn.

[301] Saxena, D. and Guha, S. “Conducting Participatory Design to Improve Algorithms in Public Services: Lessons and Challenges.” In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. Virtual Event USA: ACM, Oct. 2020, pp. 383–388. DOI: 10/gnhcvj.

[302] Schön, D. A. “Generative metaphor: A perspective on problem-setting in social policy.” In: *Metaphor and thought*. Ed. by A. Ortony. 2nd ed. Cambridge: Cambridge University Press, 1993, pp. 137–163. DOI: 10.1017/CBO9781139173865.011.

[303] Schot, J. and Rip, A. “The past and future of constructive technology assessment.” In: *Technological Forecasting and Social Change* 54.2-3 (Feb. 1997), pp. 251–268. DOI: 10/btg3nj.

[304] Scott, D. “Diversifying the Deliberative Turn: Toward an Agonistic RRI.” In: *Science, Technology, & Human Values* 48.2 (Mar. 2023), pp. 295–318. DOI: 10/gpk2pr.

[305] Selbst, A. D. and Barocas, S. “The Intuitive Appeal of Explainable Machines.” In: *SSRN Electronic Journal* (2018). DOI: 10/gdz285.

[306] Semenova, L., Rudin, C., and Parr, R. “On the Existence of Simpler Machine Learning Models.” In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul Republic of Korea: ACM, June 2022, pp. 1827–1858. DOI: 10/gr8f5d.

[307] Shaw, J. and Graham, M. “An Informational Right to the City? Code, Content, Control, and the Urbanization of Information.” In: *Antipode* 49.4 (2017), pp. 907–927. DOI: 10/gbwxcs.

[308] Shelton, T., Zook, M., and Wiig, A. “The ’actually existing smart city’.” In: *Cambridge Journal of Regions, Economy and Society* 8.1 (2015), pp. 13–25. DOI: 10/gcv8xz.

[309] Shin, D. “How do people judge the credibility of algorithmic sources?” In: *AI & SOCIETY* (Mar. 2021). DOI: 10/gk3nsh.

[310] Shneiderman, B. “Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems.” In: *ACM Transactions on Interactive Intelligent Systems* 10.4 (Oct. 2020), 26:1–26:31. DOI: 10/gh4rnz.

[311] Sinha, R. and Swearingen, K. “The role of transparency in recommender systems.” In: *CHI Extended Abstracts*. Apr. 2002. DOI: 10/bbzvjd.

[312] Sio, F. and Hoven, J. van den. “Meaningful human control over autonomous systems: A philosophical account.” In: *Frontiers Robotics AI* 5.FEB (2018), pp. 1–14. DOI: 10/gf597h.

[313] Sloane, M., Moss, E., Awomolo, O., and Forlano, L. “Participation is not a Design Fix for Machine Learning.” In: *arXiv:2007.02423 [cs]* (Aug. 2020). URL: http://arxiv.org/abs/2007.02423 (visited on 07/20/2021).

[314] Soja, E. “The city and spatial justice.” In: *Justice spatiale/Spatial justice* 1.1 (2009), pp. 1–5.

[315] Sousa, W. G. d., Melo, E. R. P. d., Bermejo, P. H. D. S., Farias, R. A. S., and Gomes, A. O. “How and where is artificial intelligence in the public sector going? A literature review and research agenda.” In: *Government Information Quarterly* 36.4 (Oct. 2019), p. 101392. DOI: 10/gg6cpj.

[316] Srnicek, N. “Data, Compute, Labour.” In: *Digital Work in the Planetary Market*. Ed. by M. Graham and F. Ferrari. The MIT Press, May 2022, pp. 241–261.

[317] Stappers, P. J. “Prototypes as central vein for knowledge development.” In: *Prototype: Design and Craft in the 20th Century* (2013), pp. 85–98. DOI: 10/gmv79z.

[318] Stappers, P. J., Visser, F. S., and Keller, A. I. “The role of prototypes and frameworks for structuring explorations by research through design.” In: *The Routledge Companion to Design Research* 25.21 (2015), pp. 167–174.

[319] Sterling, B. “Cover story: design fiction.” In: *Interactions* 16.3 (May 2009), pp. 20–24. DOI: 10/cfx568.

[320] Stilgoe, J., Owen, R., and Macnaghten, P. “Developing a framework for responsible innovation.” In: *Research Policy* 42.9 (Nov. 2013), pp. 1568–1580. DOI: 10.1016/j.respol.2013.05.008.

[321] Stolterman, E. and Wiberg, M. “Concept-Driven Interaction Design Research.” In: *Human–Computer Interaction* 25.2 (May 2010), pp. 95–118. DOI: 10/c6qftv.

[322] Stoyanovich, J. and Howe, B. *Follow the Data! Algorithmic Transparency Starts with Data Transparency*. Dec. 2018. URL: https://edu.nl/h47k3 (visited on 09/22/2021).

[323] Strengers, Y. “Peak electricity demand and social practice theories: Reframing the role of change agents in the energy sector.” In: *Energy Policy* 44 (May 2012), pp. 226–234. DOI: 10/f3zmtm.

[324] Strengers, Y. “Smart energy in everyday life: are you designing for resource man?” In: *Interactions* 21.4 (July 2014), pp. 24–31. DOI: 10/ghfk9z.

[325] Strengers, Y. *Smart energy technologies in everyday life: smart utopia?* New York: Palgrave Macmillan, 2013.

[326] Suchman, L. *Corporate Accountability*. June 2018. URL: https://edu.nl/aqp4u (visited on 08/12/2021).

[327] Suchman, L., Blomberg, J., Orr, J. E., and Trigg, R. “Reconstructing Technologies as Social Practice.” In: *American Behavioral Scientist* 43.3 (Nov. 1999), pp. 392–408. DOI: 10/d6gp23.

[328] Suchman, M. C. “Managing Legitimacy: Strategic and Institutional Approaches.” In: *Academy of Management Review* 20.3 (July 1995), pp. 571–610. DOI: 10/drhb7m.

[329] Sun, T. Q. and Medaglia, R. “Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare.” In: *Government Information Quarterly* 36.2 (Apr. 2019), pp. 368–383. DOI: 10/ggc6q7.

[330] Thackara, J. *In The Bubble: Designing In A Complex World*. First Edition. Cambridge, Mass: Mit Pr, Jan. 2005.

[331] The Cities Coalition for Digital Rights. *Cities for Digital Rights*. Aug. 2021. URL: https://edu.nl/rwnmj (visited on 09/22/2021).

[332] Thoring, K., Mueller, R., and Badke-Schaub, P. “Workshops as a Research Method: Guidelines for Designing and Evaluating Artifacts Through Workshops.” In: *Hawaii International Conference on System Sciences 2020 (HICSS-53)*. Jan. 2020. URL: https://aisel.aisnet.org/hicss-53/os/design_science_research/4.

[333] Tickle, A., Andrews, R., Golea, M., and Diederich, J. “The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks.” In: *IEEE Transactions on Neural Networks* 9.6 (Nov. 1998), pp. 1057–1068. DOI: 10/btn5vv.

[334] Tonkinwise, C. “The interaction design public intellectual.” In: *Interactions* 23.3 (Apr. 2016), pp. 24–25. DOI: 10/gj8832.

[335] Tsamados, A. et al. “The ethics of algorithms: key problems and solutions.” In: *AI & SOCIETY* (Feb. 2021). DOI: 10/gkx6tg.

[336] Tseng, Y.-S. “Assemblage thinking as a methodology for studying urban AI phenomena.” In: *AI & SOCIETY* 38.3 (June 2023), pp. 1099–1110. DOI: 10/gqmc9d.

[337] Tufte, E. R. *Visual explanations: images and quantities, evidence and narrative*. Cheshire, Conn: Graphics Press, 1997.

[338] Turel, T., Joskin, D., Geerts, F., Kaathoven, E. van, and Schouwenaar, M. “Designing a Transparent Smart Charge Point.” In: *30th International Electric Vehicle Symposium (EVS30)*. Vol. 1. Stuttgart, Germany, Oct. 2017. (Visited on 09/22/2021).

[339] Umbrello, S. “Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach.” In: *Ethics and Information Technology* 23.3 (Sept. 2021), pp. 455–464. DOI: 10/gpk2pp.

[340] Umbrello, S. “Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design.” In: *Science and Engineering Ethics* 26.2 (Apr. 2020), pp. 575–595. DOI: 10/ghc9mf.

[341] Vaccaro, K., Karahalios, K., Mulligan, D. K., Kluttz, D., and Hirsch, T. “Contestability in Algorithmic Systems.” In: *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. Austin TX USA: ACM, Nov. 2019, pp. 523–527. DOI: 10/gjr4r5.

[342] Vaccaro, K., Sandvig, C., and Karahalios, K. “"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.CSCW2 (Oct. 2020), 167:1–167:22. DOI: 10/gj7sk7.

[343] Vaccaro, K., Xiao, Z., Hamilton, K., and Karahalios, K. “Contestability For Content Moderation.” In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), pp. 1–28. DOI: 10/gnct2z.

[344] Vakarelov, O. and Rogerson, K. “The Transparency Game: Government Information, Access, and Actionability.” In: *Philosophy and Technology* 33.1 (2020), pp. 71–92. DOI: 10/ghbr9v.

[345] Van Bouwel, J. and Van Oudheusden, M. “Participation Beyond Consensus? Technology Assessments, Consensus Conferences and Democratic Modulation.” In: *Social Epistemology* 31.6 (Nov. 2017), pp. 497–513. DOI: 10/gmdmj2.

[346] Veale, M., Van Kleek, M., and Binns, R. “Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making.” In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC Canada: ACM, Apr. 2018, pp. 1–14. DOI: 10/ct4s.

[347] Verbeek, P.-P. “Beyond interaction: a short introduction to mediation theory.” In: *Interactions* 22.3 (Apr. 2015), pp. 26–31. DOI: 10/gc8zs2.

[348] Verbeek, P.-P. “Materializing Morality.” In: *Science, Technology, & Human Values* 31.3 (2006), pp. 361–380. DOI: 10/bnt8nx.

[349] Verdiesen, I., Santoni de Sio, F., and Dignum, V. “Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight.” In: *Minds and Machines* 31.1 (Mar. 2021), pp. 137–163. DOI: 10/gjd5dd.

[350] Verma, H. et al. “Rethinking the Role of AI with Physicians in Oncology: Revealing Perspectives from Clinical and Research Workflows.” In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, pp. 1–19. DOI: 10/gr869f.

[351] Vinsel, L. *You’re Doing It Wrong: Notes on Criticism and Technology Hype*. Feb. 2021. URL: https://edu.nl/74ky7 (visited on 11/09/2023).

[352] Wachter, S., Mittelstadt, B., and Floridi, L. “Why a right to explanation of automated decision-making does not exist in the general data protection regulation.” In: *International Data Privacy Law* 7.2 (2017), pp. 76–99. DOI: 10/gfc7bb.

[353] Waldman, A. “Power, Process, and Automated Decision-Making.” In: *Fordham Law Review* 88.2 (Nov. 2019), p. 613. URL: https://ir.lawnet.fordham.edu/flr/vol88/iss2/9.

[354] Walmsley, J. “Artificial intelligence and the value of transparency.” In: *AI & SOCIETY* 36.2 (June 2021), pp. 585–595. DOI: 10/gkbk5d.

[355] Wang, Q., Liu, X., Du, J., and Kong, F. “Smart Charging for Electric Vehicles: A Survey From the Algorithmic Perspective.” In: *IEEE Communications Surveys & Tutorials* 18.2 (2016), pp. 1500–1517. DOI: 10/gf6k2p.

[356] Ware, C. *Building stories*. New York: Pantheon Books, a division of Random House, Inc, 2012.

[357] Wensveen, S. and Matthews, B. “Prototypes and prototyping in design research.” In: *The Routledge Companion to Design Research* (2015), pp. 262–276.

[358] Williams, J., Fam, D., and Lopes, A. M. “Creating knowledge: visual communication design research in transdisciplinary projects.” In: *Transdisciplinary Research and Practice for Sustainability Outcomes*. Routledge, 2016.

[359] Wiltshire, G. and Ronkainen, N. “A realist approach to thematic analysis: making sense of qualitative data through experiential, inferential and dispositional themes.” In: *Journal of Critical Realism* 20.2 (Mar. 2021), pp. 159–180. DOI: 10/gf2t.

[360] Winner, L. “Do artifacts have politics?” In: *Daedalus* 109.1 (1980), pp. 121–136.

[361] Wirtz, B. W., Weyerer, J. C., and Geyer, C. “Artificial Intelligence and the Public Sector—Applications and Challenges.” In: *International Journal of Public Administration* 42.7 (May 2019), pp. 596–615. DOI: 10/gd53j4.

[362] Wyatt, A. and Galliott, J. "An Empirical Examination of the Impact of Cross-Cultural Perspectives on Value Sensitive Design for Autonomous Systems." In: *Information* 12.12 (Dec. 2021), p. 527. DOI: 10/gps2fh.

[363] Yigitcanlar, T., Agdas, D., and Degirmenci, K. "Artificial intelligence in local governments: perceptions of city managers on prospects, constraints and choices." In: *AI & SOCIETY* 38.3 (June 2023), pp. 1135–1150. DOI: 10/gqmc9f.

[364] Yildirim, N., Pushkarna, M., Goyal, N., Wattenberg, M., and Viégas, F. "Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, pp. 1–13. DOI: 10/gr6q25.

[365] Young, M. M., Bullock, J. B., and Lecy, J. D. "Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration." In: *Perspectives on Public Management and Governance* 2.4 (Nov. 2019), pp. 301–313. DOI: 10/ghs357.

[366] Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N., and 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*. Hamburg Germany: ACM, Apr. 2023, pp. 1–21. DOI: 10/gr6q36.

[367] Zajko, M. "Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates." In: *Sociology Compass* 16.3 (2022), e12962. DOI: 10/grghs7.

[368] Zeng, J., Ustun, B., and Rudin, C. "Interpretable classification models for recidivism prediction." In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 180.3 (2017), pp. 689–722. DOI: 10/gc5vs4.

[369] Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" In: *Philosophy and Technology* 32.4 (2019), pp. 661–683. DOI: 10/gd6fg2.

[370] Zhang, X., Liu, X., and Jiang, H. "A Hybrid Approach to License Plate Segmentation under Complex Conditions." In: *Third International Conference on Natural Computation (ICNC 2007)*. Haikou, China: IEEE, 2007, pp. 68–73. DOI: 10/djp92v.

[371] Zhang, Y. and Chen, X. "Explainable Recommendation: A Survey and New Perspectives." In: *Foundations and Trends® in Information Retrieval* 14.1 (2020), pp. 1–101. DOI: 10/gg7q3s.

[372] Zuiderwijk, A., Chen, Y.-C., and Salem, F. “Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda.” In: *Government Information Quarterly* 38.3 (July 2021), p. 101577. DOI: 10/gjzxk5.

[373] Zuo, C., Liang, K., Jiang, Z. L., Shao, J., and Fang, J. “Cost-effective privacy-preserving vehicular urban sensing system.” In: *Personal and Ubiquitous Computing* 21.5 (Oct. 2017), pp. 893–901. DOI: 10/gb4ph6.

Summary

This thesis is about using artificial intelligence (AI) in the public sector as part of policy execution. The research's main concern is that public AI can harm people's autonomy. To address this problem, we should ensure that system actions do not unduly limit citizens' control over their lives. We should also give citizens a say in when, how, and for what purpose those systems are developed and used. The author calls the system quality that enables this form of control *contestability*. The research question central to this thesis is: *What socio-technical design interventions enhance the contestability of public AI systems?*

The approach is framed as constructive design research. Researchers collaborate with design practitioners to create artifacts and use them as 'instruments' to generate data. The design methods employed include interaction design (prototyping), speculative design (animation), and information design (visual explanations). The actual real-world cases that are used to ground the research in particular contexts are smart electric vehicle charging, camera cars used for urban monitoring, and fraud detection risk models. The empirical work was done in Amsterdam in close collaboration with the city government, technology companies, and design firms.

The findings include an account of the differences between how citizens and experts think about AI transparency, a provisional framework that describes contestable AI features and practices, a description of the main challenges facing the implementation of contestable AI in local government practice, and an exploration of the guiding concepts (metaphors) designers use when considering public AI.

The research has implications for responsible, explainable, and human-centered AI, civic participation in public AI, and design research and practice.

With regards to human-centered AI, trust results from systems that operate safely and reliably, allowing citizens to keep control and enabling subjects to quickly report and recover from errors. Rather than focus on AI opacity in isolation, this research indicates we should focus on how AI impacts people's everyday concerns.

Concerning civic participation, procedures for handling citizens' reports, questions, complaints, objections, and appeals should not be considered a mere matter of legal compliance. This research shows that they should be seen as feedback mechanisms for continuous system improvement. Such procedures should not be tacked on but deeply integrated with the primary systems. The thesis demonstrates how a crucial element of these procedures is a two-way dialogue between subjects and controllers on equal footing. Although there is a role for direct democracy in ensuring civic control of public AI, existing representative democratic institutions should be revitalized and connected more strongly with public AI development.

With regards to design, to ensure that organizations adopt successful public AI pilots, this research suggests designers should act as stewards who guide those systems from inception through implementation. Speculative design can be used for more than merely asking questions. It can also produce instruments that generate data that help answer them. The distinction with traditional design is that the types of questions you can ask with it and the answers you will get are different: focused more on relations, context, and time. Design knowledge is most useful for practitioners when it offers guidance for ideating, form-giving, and specifying in alignment with each other.

The author concludes with a call for design researchers and practitioners in AI and beyond to become more aware of and engage with political philosophy to understand better how their work supports or undermines particular models of democracy.

Samenvatting

Dit proefschrift gaat over het gebruik van artificiële intelligentie (AI) in de publieke sector als onderdeel van de beleidsuitvoering. De voornaamste zorg van het onderzoek is dat publieke AI de autonomie van mensen kan schaden. Om dit probleem aan te pakken moeten we ervoor zorgen dat systeemacties de controle van burgers over hun leven niet ongewenst beperken. We moeten burgers ook inspraak geven over wanneer, hoe en voor welk doel deze systemen worden ontwikkeld en gebruikt. De auteur noemt de systeemkwaliteit die deze vorm van controle mogelijk maakt *contestability* (betwistbaarheid). De onderzoeksvraag die centraal staat in dit proefschrift is: *Welke sociaal-technische ontwerpinterventies vergroten de betwistbaarheid van publieke AI-systemen?*

De aanpak wordt beschreven als constructief ontwerponderzoek. Onderzoekers werken samen met ontwerpprofessionals om artefacten te creëren en deze te gebruiken als 'instrumenten' om gegevens te genereren. De gebruikte ontwerpmethoden omvatten interactieontwerp (prototyping), speculatief ontwerp (animatie) en informatieontwerp (visuele uitleg). De praktijkvoorbeelden die worden gebruikt om het onderzoek in bepaalde contexten te gronden zijn het slim opladen van elektrische voertuigen, camera-auto's die worden gebruikt voor monitoring van de publieke ruimte en risicomodellen voor fraudedetectie. Het empirische werk werd in Amsterdam gedaan in nauwe samenwerking met het stadsbestuur, technologiebedrijven en ontwerpbureaus.

De bevindingen omvatten een verslag van de verschillen tussen hoe burgers en experts denken over AI-transparantie, een voorlopig ontwerpkader dat functies en praktijken die bijdragen aan betwistbaarheid beschrijft, een beschrijving van de belangrijkste uitdagingen waarmee de implementatie van betwistbare AI in de praktijk van lokale overheden wordt geconfronteerd, en een verkenning van de leidende concepten (metaforen) die ontwerpers gebruiken bij het overwegen van publieke AI.

Het onderzoek heeft implicaties voor verantwoorde, verklaarbare en mensgerichte AI, burgerparticipatie in publieke AI, en ontwerponderzoek en -praktijk.

Met betrekking tot mensgerichte AI stelt de auteur dat vertrouwen voortvloeit uit systemen die veilig en betrouwbaar werken, waardoor burgers de controle kunnen behouden en snel fouten kunnen rapporteren en herstellen. In plaats van ons te concentreren op de ondoorzichtigheid van AI op zichzelf, geeft dit onderzoek aan dat we ons moeten richten op de manier waarop AI de dagelijkse zorgen van mensen raakt.

Wat burgerparticipatie betreft, mogen de procedures voor de behandeling van meldingen, vragen, klachten, bezwaren en beroepen van burgers niet louter als een kwestie van wettelijke naleving worden beschouwd. Uit dit onderzoek blijkt dat ze gezien moeten worden als feedbackmechanismen voor continue systeemverbetering. Dergelijke procedures moeten niet worden vastgeplakt, maar diep worden geïntegreerd met de primaire systemen. Het proefschrift laat zien hoe een cruciaal element van deze procedures een tweerichtingsdialoog is tussen proefpersonen en controleurs op gelijke voet. Hoewel er een rol is weggelegd voor de directe democratie bij het waarborgen van de civiele controle over publieke AI, moeten bestaande representatieve democratische instellingen nieuw leven worden ingeblazen en sterker worden verbonden met de ontwikkeling van publieke AI.

Met betrekking tot ontwerp, om ervoor te zorgen dat organisaties succesvolle publieke AI-pilots adopteren, suggereert dit onderzoek dat ontwerpers moeten optreden als 'opzichters' die deze systemen vanaf het begin tot aan de implementatie begeleiden. Speculatief ontwerp kan voor meer worden gebruikt dan alleen het stellen van vragen. Het kan ook instrumenten produceren die gegevens genereren die helpen vragen te beantwoorden. Het verschil met traditioneel ontwerp is dat het soort vragen dat we ermee kunnen stellen en de antwoorden die we krijgen anders zijn: meer gericht op relaties, context en tijd. Ontwerpkennis is het nuttigst voor mensen uit de praktijk als deze een leidraad biedt voor het bedenken, vormgeven en specificeren in samenhang met elkaar.

De auteur sluit af met een oproep aan ontwerponderzoekers en praktijkmensen op het gebied van AI en daarbuiten om zich meer bezig te houden met politieke filosofie om beter te begrijpen hoe hun werk bepaalde modellen van democratie ondersteunt of ondermijnt.

Acknowledgments

It is Saturday, January 16, 2016. I am sitting outside Jewel Coffee on Singapore's Orchard Road, thinking about what to do next with my career, taking notes. I want to dig deeper into the impact of AI on design. But where best to do this? Through a process of elimination, I end up with one remaining option: pursue a PhD. And go to Delft to do it. A day later, in my diary, I wrote, "It is really interesting. And exciting. I hope this will work out."

Eight years, four months, and seven days later, I will defend this thesis. It took a while, but we got there. The journey indeed turned out to be interesting. And exciting. Things worked out in the end, but only because I had help from many generous people. It is impossible to list all of them, and my gratitude is far greater than can be expressed here. What follows will have to do.

First of all, my supervisors.

My thanks to Gerd for taking that initial gamble on me. Also, credit where credit is due, it was Gerd who proposed "contestability" as an encapsulation of my interests. His relentless focus on those twin scientific values of novelty and relevance pushed me to ever-greater heights. For some reason, Gerd saw potential in me and would only be satisfied once it was realized.

To Neelke, for her steady hand. She would always make time for me when needed and expressed sincere interest in the person behind the researcher. Neelke has a surgical eye for detail. I learned a lot from her about how to construct an argument and the importance of distinguishing between the essential and the extraneous.

Finally, I thank Ianus for bringing me into TU Delft and helping me make those early connections, to PJ in particular, that would ultimately lead to this project. Later, he agreed to become my daily supervisor, ensuring I stayed true to my designer roots. Ianus always joked that I should simultaneously do a PhD project and become a dad. I hope he's happy now.

Next, I want to give thanks to people who played a pivotal role in the early stages of my time at IDE.

To PJ, for those early talks, exploring my research interests. And also, crucially, for introducing me to Gerd. I am grateful to PJ for the guidance and inspiration regarding research through design.

To Péter and Holly, for being my collaborators when I had yet to find my place at IDE, giving me insight into what it would be like to do a PhD. Péter, in particular, for those early explorations of AI and design we did together that were a lot of fun but also helped me sharpen my research plans.

And to Sacha for being such a kind and welcoming mentor in the early stages of my PhD and helping me further shape my supervisory team to be the best it could be.

Thanks are also due to my wonderful colleagues in the Knowledge & Intelligence Design (KInD) section: Achilleas, Adrie, Alejandra, Alessandro, Anelia, Carlo, Céline, Denis, Di, Evangelos, Francesca, Garoa, Himanshu, Hosana, Jacky, James, Jeff, Ludovica, Mireia, Peter, Roos, Ruben, Samuel, Sara, Sergei, Shatha, Tianhao, Tilman, Uğur, Vasilis, Wilfred, Wo, and Yunzhong.

My thanks to Jacky, particularly, for being such a wonderful roommate in those early days before there really even was a KInD section to speak of. Jacky patiently answered my ignorant questions and listened to my many rants, providing much-needed camaraderie.

To Denis and Mireia for generously collaborating with me on the article that makes up this thesis' final chapter. Thanks to Denis for his support while running the workshops. To Mireia, thanks for helping me think through early outlines and later helping me make sense of the analysis later. It has been wonderful to find a fellow young researcher in Mireia with just as much, if not more, passion for the topic of contestable AI. I am also very grateful to Mireia for agreeing to be my paranymph.

Thanks to the KInD PhDs—for the many coffees, lunches, and informal chats.

And to Alessandro for the stimulating discussions about science, politics, and everything in between.

A thank you to all the students who joined the AI & Society master electives over the years. Developing and teaching the elective significantly catalyzed my thinking about contestability. I could not have done it without their interest and engagement.

Also, a big thanks to those IDE master graduation students to whom I had the privilege of being a mentor. It is always energizing and exciting to accom-

pany young designers on this final part of their journey. Thanks to Fabian and Laura, in particular, for taking elements from my research as a starting point and spinning them out into exciting and attractive designs.

Much of the research reported on in this thesis was only possible with the kind participation of many, many people. I sincerely thank them for their time and patience. All the EV drivers who participated in the Transparent Charging Station prototype field test. All the experts who helped assess the Contestable Camera Cars concept video and the City of Amsterdam civil servants who participated in the related interview study. All the students who joined the trial workshop of the Envisioning Contestability Loops study and the designers who participated in the related design workshops.

Within TU Delft, I thank Roy for the many stimulating conversations related to speculative and critical design. To Roel, thank you for our ongoing chats about AI safety governance and politics. Also, a shout-out to Sem, whose PhD research has been converging on issues similar to mine and who has become a valuable sparring partner.

More thanks to collaborators further afield are also in order.

At the BRIDE project, I thank my University of Twente colleagues Michael and Sage for the enjoyable and inspiring collaborations. Sage, in particular, for all our chats over the years about our PhDs, comparing notes, letting off steam, and sharing ideas.

At the Amsterdam Institute for Advanced Metropolitan Studies and the Responsible Sensing Lab, my thanks go out in particular to Thijs for being that critical link between myself and "the field." Thijs was always interested in hearing my ideas, was instrumental in bringing me into projects, and was always willing to support my studies by connecting me to people, acquiring funding, and more.

Also, I thank Frank and Lotte at ElaadNL for their inspiration and support during the 'UI for Smart EV Charging' project. A special thanks to Lotte, in particular, for the enjoyable collaboration on the planning and execution of the field studies.

Those design artifacts that illustrate this work would not have been half as appealing if it weren't for the valuable efforts of my design specialist collaborators. At The Incredible Machine, Harm and Marcel and their team for the work on the Transparent Charging Station prototype. At Trim Tab Pictures, my thanks to Simon for the creation of the Contestable Camera Cars concept video.

Leon de Korte, for his work creating the Contestability Loops infographic. And finally, Joost Stokhof, for the art that adorns the cover and interior of this thesis.

At the Amsterdam University of Applied Sciences, I thank my collaborators in the Civic Interaction Design group: Jorgen, Martijn, Mike, and Tessa. Our work together has become a fruitful space for further exploring the implications of my research for design practice and smart city governance.

At Umeå University, my thanks to fellow contestability PhD researcher Rob for the stimulating conversations. Seeing a completely different take on the same concept has been interesting.

Then, there are the many technology design and ethics thinkers who, in part, inspired me to pursue this PhD and were kind enough to spend some of their time discussing the topic, reviewing materials, and offering feedback: Cennydd, Claudio, Hans, Jussi, Marco, Miguel, and Sebastian, my thanks to all of them.

To finish, some personal notes of gratitude.

The Tabletop Fridays crew: Dennis, Hessel, Joost H, Joost W, Michiel, Niels, and Thomas, for those many, many, *many* evenings of playing board games and D&D together, online and in-person. I am pretty sure I would have gone insane if it had not been for this pressure valve. A special thanks to Joost H for agreeing to be my paranymph.

To my parents, Nicolette and Ben, for all their love, care, attention, and encouragement. My mother, for always showing an interest in my work and instilling that strong sense of justice in me. My father, for passing on his commitment to quality and putting me on my interaction design journey in those heady late-nineties days. I may not have inherited his knack for mathematics, but acquiring a PhD from his alma mater hopefully makes up for that a little bit.

My sister and brother, Maggie and Ties, for all the companionship and support over the years through thick and thin. And now, together with my brother and sister-in-law, Sybren and Robin, for sharing the joys of each of us starting families and marveling at each others' brood. Thanks to Ties for supporting me in getting this thesis' layout and typography up to scratch and making its printing 'extra nice.'

My parents-in-law, Jeanne and Martin, for their wisdom, kindness, and care. And my sister-in-law, Jacintha, for her cheerfulness and curiosity. The second family they have allowed me to become a part of is a true oasis in these hurried times.

To my sons, Seth and Jonas. Born roughly four months before I started my PhD, these two guys have ensured I never lose sight of the most essential things in life.

I may have learned a few new tricks these past years, but they pale compared to what these two have already accomplished. I could not be more proud and cannot wait to see what they do next.

And finally, to my wife, Lieke, for her unwavering support, encouragement, and love. Not a day goes by that I do not consider myself the luckiest bastard on this planet for sharing this life with her.

About the Author

Christiaan Pieter (Kars) Alfrink was born on April 12, 1980 in Blokzijl, the Netherlands.

From 1992 to 1998, he attended St. Bonifatiuscollege Utrecht, where he completed his pre-university education (VWO). From 1998 to 2002, he attended Utrecht School of the Arts, where he acquired a Bachelor of Arts in Interaction Design and a European Media Master of Arts in Gaming (cum laude), validated by the University of Portsmouth.

Over the following 15+ years, Kars was active as an interaction design practitioner, public speaker, reluctant entrepreneur, and design community organizer. Kars began his career as an interaction designer at several web agencies. He also worked as a teacher and researcher in higher arts education. Some subsequent highlights include: initiating and co-curating the Dutch offshoot of This Happened, a series of events about the stories behind interaction design (`thishappened.nl`); from 2009 to 2016 founding and acting as a partner at Hubbub (`hubbub.eu`), a boutique playful design agency; and cofounding and coordinating Tech Solidarity NL (`techsolidarity.nl`), a grassroots community of Dutch tech workers who seek to advance the design and development of more just and egalitarian technology.

Kars conducted this research from 2018 to 2023 at Delft University of Technology in the Sustainable Design Engineering department of the faculty of Industrial Design Engineering.

In his free time, Kars enjoys refereeing an old-school Dungeons & Dragons roleplaying campaign, collecting pulp science-fantasy novels and comics, and cooking Southeast Asian dishes to stave off *fernweh*. He lives in Utrecht, the Netherlands, with his wife Lieke and their two sons, Seth and Jonas.

List of Publications

Refereed Journal Articles and Conference Papers

- Alfrink, K., Keller, I., Yurrita, M., Bulygin, D., Kortuem, G., and Doorn, N. "Envisioning Contestability Loops: Evaluating the Agonistic Arena as a Generative Metaphor for Public AI." in: *She Ji: The Journal of Design, Economics, and Innovation* (in press)
- Alfrink, K., Keller, I., Doorn, N., and Kortuem, G. "Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, pp. 1–16. DOI: 10/gr5wcx 🏆
- Alfrink, K., Keller, I., Kortuem, G., and Doorn, N. "Contestable AI by Design: Towards a Framework." In: *Minds and Machines* 33.4 (Aug. 2022), pp. 613–639. DOI: 10/gqnjcs
- Alfrink, K., Keller, I., Doorn, N., and Kortuem, G. "Tensions in Transparent Urban AI: Designing a Smart Electric Vehicle Charge Point." In: *AI & Society* 38.3 (Mar. 2022), pp. 1049–1065. DOI: 10/gpszwh

Conference Workshop Papers

- Alfrink, K., Turel, T., Keller, I., Doorn, N., and Kortuem, G. "Contestable City Algorithms." In: *Participatory Approaches to Machine Learning*. Ed. by B. Kulynych, D. Madras, S. Milli, I. D. Raji, A. Zhou, and R. Zemel. Workshop at the International Conference on Machine Learning. July 2020. URL: https://participatoryml.github.io PDF: edu.nl/m8xc8
- Alfrink, K., Turel, T., and Kortuem, G. "Designing a Smart Electric Vehicle Charge Point for Algorithmic Transparency: Doing Harm by Doing Good?" In: *Urban AI: Formulating an Agenda for the Interdisciplinary Research of Artificial Intelligence in Cities*. Ed. by A. Luusua and J. Ylipulli. Workshop at the

Designing Interactive Systems Conference. July 2020. DOI: 10/gjr4r6 PDF: edu.nl/uvddg

- Alfrink, K., Turel, T., and Kortuem, G. “Contestable City Algorithms.” In: *Contestability in Algorithmic Systems*. Ed. by K. Vaccaro, K. Karahalios, D. K. Mulligan, D. Kluttz, and T. Hirsch. Workshop at the Conference on Computer Supported Cooperative Work and Social Computing. Nov. 2019. DOI: 10/gjr4r5 PDF: edu.nl/8a6uu

Appendix A
Contestable AI by Design: Summary of Reviewed Literature

Table A.1
Included sources and their related features and practices.

Source	Features	Practices
Almada [9]	built-in safeguards; explanations; intervention requests; tools for scrutiny	agonistic approaches; ex-ante safeguards; QA after deploy; QA during dev
Aler Tubella et al. [4]	explanations; tools for scrutiny	ex-ante safeguards; QA after deploy
Bayamlıoğlu [22]	explanations; interactive control; intervention requests; tools for scrutiny	ex-ante safeguards; QA after deploy; 3rd party oversight
Brkan [47]	explanations; intervention requests	ex-ante safeguards
Crawford [66]	explanations	3rd party oversight
Edwards and Veale [91]	explanations; intervention requests	ex-ante safeguards; 3rd party oversight
Elkin-Koren [95]	built-in safeguards; intervention requests	QA during dev; 3rd party oversight
Henin and Le Métayer [146]	explanations; intervention requests; tools for scrutiny	agonistic approaches; ex-ante safeguards
Hirsch et al. [153]	explanations; interactive control; intervention requests; tools for scrutiny	ex-ante safeguards; QA after deploy; QA during dev; risk mitigation
Jewell [172]	interactive control	–
Kariotis and J. Mir [179]	tools for scrutiny	agonistic approaches; ex-ante safeguards; QA during dev
König and Wenzelburger [192]	agonistic approaches	–

Table A.1
Included sources and their related features and practices (continued).

Source	Features	Practices
Lyons et al. [223]	explanations; intervention requests; tools for scrutiny	ex-ante safeguards; risk mitigation; 3rd party oversight;
Ploug and Holm [278]	explanations; intervention requests	QA during dev; risk mitigation
Sarra [298]	explanations; intervention requests	ex-ante safeguards
Vaccaro et al. [341]	explanations; interactive control; intervention requests; tools for scrutiny	agonistic approaches; QA during dev; risk mitigation; 3rd party oversight
Vaccaro et al. [342]	interactive control; intervention requests; tools for scrutiny	QA after deploy; QA during dev; risk mitigation; 3rd party oversight
Vaccaro et al. [343]	explanations; intervention requests	agonistic approaches; QA after deploy
Walmsley [354]	intervention requests	ex-ante safeguards; QA during dev; QA after deploy

Table A.2
Features and their related sources.

Feature	Sources
Built-in safeguards against harmful behavior	Almada [9] and Elkin-Koren [95]
Interactive control over automated decisions	Bayamlıoğlu [22], Hirsch et al. [153], Jewell [172], and Vaccaro et al. [341, 342]
Explanations of system behavior	Aler Tubella et al. [4], Almada [9], Bayamlıoğlu [22], Brkan [47], Crawford [66], Edwards and Veale [91], Henin and Le Métayer [146], Hirsch et al. [153], Lyons et al. [223], Ploug and Holm [278], Sarra [298], and Vaccaro et al. [341, 343]
Human review and intervention requests	Almada [9], Bayamlıoğlu [22], Brkan [47], Edwards and Veale [91], Elkin-Koren [95], Henin and Le Métayer [146], Hirsch et al. [153], Lyons et al. [223], Ploug and Holm [278], Sarra [298], Vaccaro et al. [341–343], and Walmsley [354]

Table A.2
Features and their related sources (continued).

Feature	Sources
Tools for scrutiny by subjects or third parties	Aler Tubella et al. [4], Almada [9], Bayamlıoğlu [22], Henin and Le Métayer [146], Hirsch et al. [153], Kariotis and J. Mir [179], Lyons et al. [223], and Vaccaro et al. [341, 342]

Table A.3
Practices and their related sources.

Practice	Sources
Ex-ante safeguards	Aler Tubella et al. [4], Almada [9], Bayamlıoğlu [22], Brkan [47], Edwards and Veale [91], Henin and Le Métayer [146], Hirsch et al. [153], Kariotis and J. Mir [179], Lyons et al. [223], Sarra [298], and Walmsley [354]
Agonistic approaches to ML development	Almada [9], Henin and Le Métayer [146], Kariotis and J. Mir [179], König and Wenzelburger [192], and Vaccaro et al. [341, 343]
Quality assurance during development	Almada [9], Elkin-Koren [95], Hirsch et al. [153], Kariotis and J. Mir [179], Ploug and Holm [278], Vaccaro et al. [341, 342], and Walmsley [354]
Quality assurance after deployment	Aler Tubella et al. [4], Almada [9], Bayamlıoğlu [22], Hirsch et al. [153], Vaccaro et al. [342, 343], and Walmsley [354]
Risk mitigation strategies	Hirsch et al. [153], Lyons et al. [223], Ploug and Holm [278], and Vaccaro et al. [341, 342]
Third-party oversight	Bayamlıoğlu [22], Crawford [66], Edwards and Veale [91], Elkin-Koren [95], Lyons et al. [223], and Vaccaro et al. [341, 342]

Appendix B
Envisioning Contestability Loops: Creative Brief

Motivation

AI is increasingly used by public sector actors to support, augment and automate decision-making. Such systems lack democratic legitimacy. This can be improved by ensuring systems are *contestable*: Open and responsive to human intervention throughout their lifecycle, establishing a dialectical relationship between decision subjects and system operators.

Objective

The aim of this project is to create a **visual explanation** that enables professional interaction designers to create **concept designs** for **public AI** systems that are **contestable**. The contents of this visual explanation are derived from the *features* section of the "Contestable AI by Design"-framework (Figure 3.2) [7], and the "five contestability loops"- model (Figure 4.2) [5].[1] The envisioned use case of the visual explanation is that it serves as a source of guidance and inspiration for design practitioners in the early stages of design projects dealing with public AI systems. The visual explanation will be evaluated as part of a scientific study, using a half-day workshop in which designers are tasked with improving the contestability of a real-world public AI system that is presented to them by a representative of the municipal (city) government that owns and operates it.

Success criteria

The visual explanation should be…

1. The original brief included the figure images. These are omitted here to avoid duplication.

1. **Effective, useful:** Supports the intended task. Concept designs created with it share properties with those described in the "Contestable AI by Design"-framework.
2. **Learnable:** Easy to understand.
3. **Efficient:** Quick to use.
4. **Complete, self-contained:** Contains all necessary information.
5. **Flexible:** Adaptable to individual designers' preferred way of working.
6. **Generative:** Inspires novel ideas for future designs that do not simply mirror what is represented in the visual explanation or underlying framework. Has the correct level of abstraction.
7. **Delightful, attractive:** Pleasant to use and attractive to perceive.

Deliverables

The ultimate deliverable of the project is (at minimum) a single visual explanation that can be printed by designers in their studios, or easily viewed on a single display or projector while working. This implies a print size of up to A3, and a landscape orientation.

Note: It is unlikely that a single static A3-size image will be insufficient to convey all that we need to. This will have to be further determined at design time.

Content

The visual explanation should convey the following content:

- Actors:
  - Citizens (a.k.a. "decision subjects")
  - Developers (perhaps distinguishing between internal and external ones)
  - Policy-makers (alderpersons, mayor, council members, …)
  - AI system (data inputs, models, output predictions)
  - Civil servant (a.k.a. "human controller")
  - "Third parties" (e.g., external oversight bodies)
- Contestability features:
  - Interactive controls – for civil servants
  - Explanations (*justifications*) – for citizens
  - Intervention requests (appeals) – channels for voice, arenas for debate, obligation to review/respond/reconsider decision
  - Tools for scrutiny – for citizens, third parties

  - (These are all taken from the framework. We leave out Built-In Safeguards because it is not central to contestations.)
- Loops:
  - Appeals (loop L1). These map onto intervention requests, above.
  - Monitoring of decision-appeals loops leading back to development (L4) and policy-making (L5). These *do* imply technical system features that are not captured by the original framework.
  - (We choose to leave out Participation in Development (L2) and in Policy-Making (L3) because these are practice-related.)

Additional requirements

- Dynamism, temporality: The visual explanation should show how a system, under the influence of contestations, shifts from a present state to a future state.
- Context: We need one or more real-world examples, or conceptual metaphors [202], that we can use to visually communicate the abstract features and loops described above.

Appearance

Some example visual explanations that serve as starting points for visual design.[2]

Waterwerk by Carlijn Kingma is a visual explanation of the contemporary monetary system that uses water as its central conceptual metaphor. It will be necessary for us to deploy metaphor as well. Picking the right one will require some careful consideration.[3]

Building Stories by Chris Ware is strictly speaking not a visual explanation, but a (non-linear) comic. I like the way Ware mixes architecture and narrative in one image, and the reader can start anywhere. Relevant because we will be explaining interactions that happen between actors over time, that are in turn distributed in time and space [356].

The State of the Beaches is one of many infographics by Megan Jaegerman that are described by Edward Tufte as "some of the best news graphics ever,"[4]

2. The original brief included images of the works mentioned. These are not reproduced here for copyright reasons.
3. `https://www.waterworksofmoney.com`
4. `https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0002w4`

because she mixes wit with elegance and informativeness. I like how this image has the coastline as a backbone and then various callouts that zoom into particular aspects.

Definitions

- **Visual explanation:** "*Pictures of verbs*, the representation of mechanism and motion, of process and dynamics, of causes and effects, of explanation and narrative"[337]. In our case we use visual explanations as a form of intermediate-level design knowledge – i.e., somewhere between particular design instances, and general theory [161].
- **Design concept:** Portrayals of future designs [321]. As opposed to design *artifacts*.
- **Artificial intelligence (AI):** "[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"[326].
- **Public AI:** AI used by public sector actors for supporting, augmenting or automating decisions [265].
- **Contestability:** Open and responsive to human intervention, throughout the system lifecycle, establishing a dialectical relationship between decision subjects and system operators [7].

Appendix C
Envisioning Contestability Loops: Infographic Description

The infographic shows a generic public AI system. It also shows several mechanisms that can be added to create contestability loops. We walk through each in turn.

First, we have a schematic public *human-AI system* (Figure C.1). We are taking a socio-technical view [327]. The 'system' consists not only of technology but also humans and their practices. This graphic presupposes that a system is already in place. It does not depict its initial design and development.

As a first step, data comes into the system. Using a model, or set of rules, the AI then uses this data to make a prediction. Then, we have one of two options: either the system fully automatically translates the prediction into a decision, or a human decides based on this prediction (and perhaps additional information). In both cases, the decision impacts a citizen significantly. We call this person the decision subject.

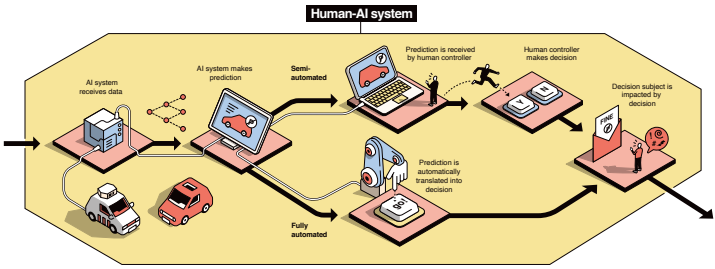


Figure C.1
Infographic detail: Human-AI system.

Now we move on to the contestability mechanisms. First, *interactive controls* (Figure C.2) intervene in the prediction-to-decision step. Humans, controllers,

or subjects may have access to additional information that the AI does not. They can supplement the prediction with this information and have it updated.

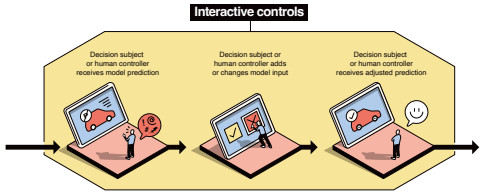


Figure C.2
Infographic detail: Interactive controls.

Next, we look at contestation after a decision has been made. So-called *intervention requests* (Figure C.3). These can be broken down into explanations, channels for voice, arenas for debate, and the obligation to respond. First, a subject needs to be provided with an explanation of how a decision was made and why it is desirable. Then, a subject must have access to channels by which they can express their objection. This appeal should lead to a dialogical exchange of viewpoints with a system representative in a so-called arena. Finally, the system operators should be obliged to respond to objections. The obligation to respond also implies that decisions must be reversible or repairable.

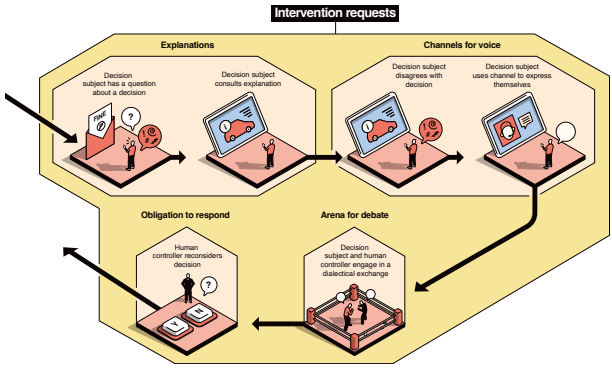


Figure C.3
Infographic detail: Intervention requests.

Connected to the previous decision-appeal loop is a second-order *monitoring* loop (Figure C.4). Here, a record of all decision appeals is kept. This record is analyzed for patterns that indicate systemic shortcomings. A human operator

is alerted to investigate if such a pattern is suspected. It is then up to the human to decide on further action. A systemic flaw can require technology revision or, further upstream, to revise policy.

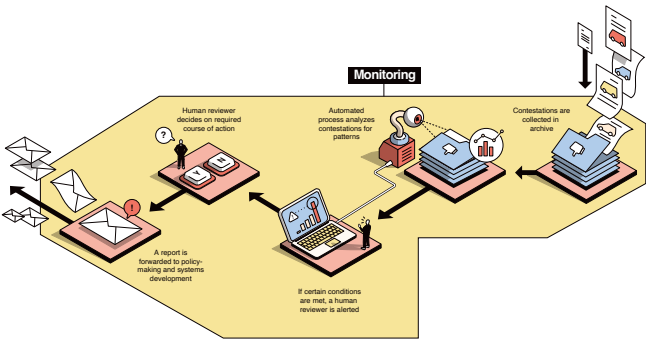


Figure C.4
Infographic detail: Monitoring.

The following mechanism is about *global* contestability. *Tools for scrutiny* (Figure C.5) are public resources that explain and justify the system as a whole. These can be used by subjects or the broad category of 'third party' actors, including journalists, and civil society organizations, to hold the system and its operators to account. This mechanism is connected to policy and system development, as well.

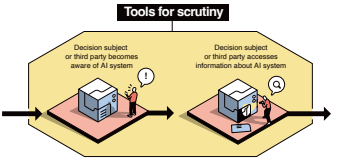


Figure C.5
Infographic detail: Tools for scrutiny.

Since we are explicitly dealing with public AI systems in this infographic, we also have a mechanism for *policy and system development* (Figure C.6). Citizens have access to various political tools for influencing systems. By means of representative democracy, they can elect representatives that shape the policies that ultimately lead to systems. However, citizens can also more directly participate in policy and technology development. This mechanism produces the policies that directly govern human controller behavior or are translated into technology.

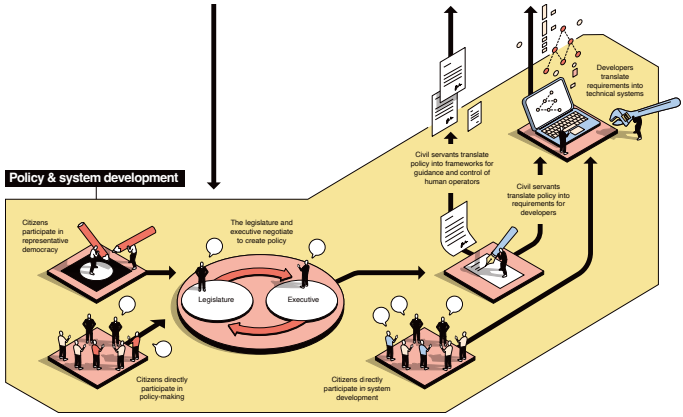


Figure C.6
Infographic detail: Policy and system development.

The flow at the bottom (Figure C.7) shows the overarching motivation for all these mechanisms. It shows how, under the influence of ongoing contestation, systems are pushed over time toward an increasingly more accountable and legitimate state.

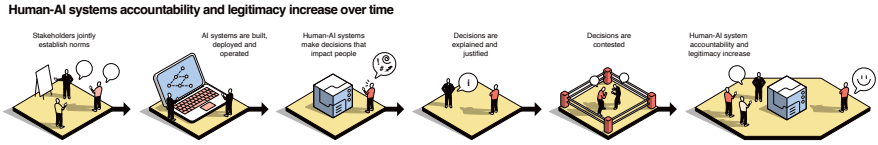


Figure C.7
Infographic detail: Accountability and legitimacy increase over time.

Appendix D

Envisioning Contestability Loops: Workshop Schedule

Total duration: 3–4 hours.

Set-up

- ☐ Private room with free wall space, tables, chairs
- ☐ Projector
- ☐ Drinks, snacks
- ☐ Handouts: visual explanation and case description
- ☐ Information sheets and informed consent forms
- ☐ Demographics forms
- ☐ Grading forms
- ☐ Presentation on laptop
- ☐ A3 paper, post-its, pencils, and markers

Walk-in

- Participants arrive

Opening (15 mins)

- Welcome
- Introductions of all participants
- Overview of the workshop schedule
- Questions at this point?
- Informed consent (human research ethics)
- Demographics form
- BREAK

Knowledge transfer (30 mins)

- Contestable AI (brief presentation)
- Visual explanation (big on screen and printed as handout)
- Case (brief presentation, and printed as handout)
- Questions?
- BREAK

Concept design work (60–90 mins)

- Design exercise explanation
- [Design work happens here.]
- BREAK mid-way and again at the end

Presentations and crit (10 mins)

- Presentations by participants of design concepts, including rationale
- Clarifying questions only
- BREAK

Focus group (30–60 mins)

- Individual completion of evaluation form
- [Focus group discussion happens here. See separate guide for details.]

Closing (5 mins)

- Final questions or comments from participants
- Thanks and closing

Wrap-up

Collect all…

☐ Signed consent forms
☐ Demographics forms
☐ Concept design results
☐ Completed grading forms

Appendix E

Envisioning Contestability Loops: Case Description

Overview

- Amsterdam has limited living space for both citizens and visitors.
- Citizens who want to rent out their home to tourists must meet certain requirements.
- A maximum of 30 nights per year and 4 people at a time is allowed, and it must be reported to the municipality.
- The municipality receives reports of possible illegal holiday rentals and investigated with the help of an algorithm pilot program starting July 2020.
- The algorithm analyzes data from related illegal housing cases of the past 5 years to calculate the probability of an illegal holiday rental situation.

Functional description

Data

Identity and housing rights data: Minimized dataset from the Personal Records Database (BRP), showing information about the identity and housing rights of the residents; specifically: name; date of birth; gender; date of residence in Amsterdam date of residence at the address; family composition; date of death.

Buildings data: Minimized dataset from the Registry of Addresses and Buildings (BAG), showing information about the building; specifically; address, street code, postal code; description of the property; Amsterdam BAG-code, national BAG-code; the type of home (rent, social rent / free sector, owner-occupied); number of rooms; floor surface area; floor number on which the front door of the apartment resides; number of building layers; description of the floor of the residential property.

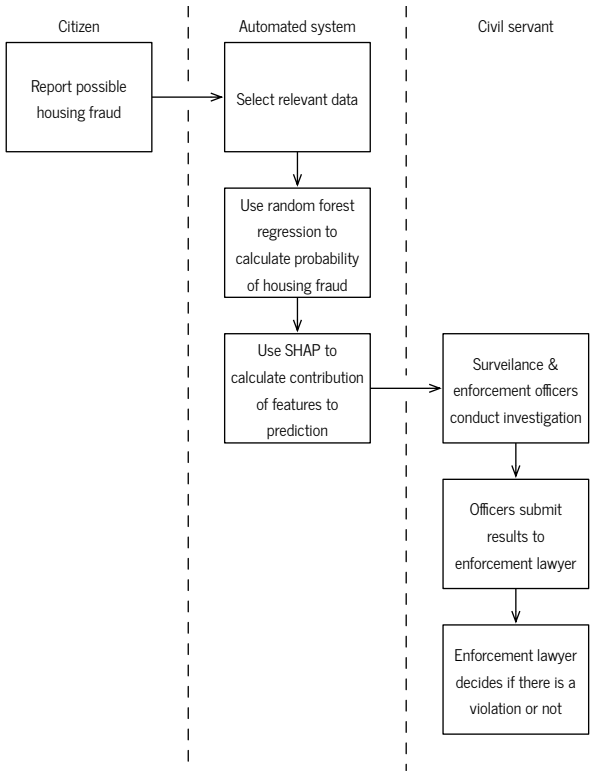


Figure E.1
Flowchart of algorithmic system used for enforcement of illegal vacation rentals. Adapted from the original by Linda van de Fliert.

Prior illegal housing cases: Data from any related illegal housing cases; specifically: starting date of investigation / report; stage of investigation; report code number; violation code number; investigator code number; anonymous reporter yes/no; situation sketch; user that created the report (including date), or edited the report (including date); handling code number (type of case, allocation to team); date when case closed; reason why case closed.

Model

- A "random forest regression" algorithm has been developed to find patterns in a large amount of information about illegal housing.

- The algorithm calculates the probability of illegal holiday rental at an address based on related illegal housing cases from the past 5 years.
- The algorithm uses the probability tree principle to perform mathematical calculations and take an average to generate the expectation of illegal holiday rental at an address.
- The "SHAP" (SHapley Additive exPlanations) method is used to explain the features in the data that resulted in high or low suspicion of illegal housing so that employees can make a well-considered decision.
- The algorithm must be carefully calibrated to avoid overfitting and categorizes continuous data points to better reach a conclusion.

Non-discrimination

- The algorithm was developed using a privacy impact assessment to ensure that sensitive information is not included.
- The dataset used in the algorithm only includes critical information to determine if the Housing Act is violated.
- The data used for the algorithm comes from previous illegal holiday rental cases to ensure good-quality data.
- The algorithm may indirectly lead to undesirable differences in treatment between cases, so the AI Fairness 360 toolkit is used to address algorithmic bias during the pilot.
- Further research will be conducted to ensure that the algorithm is fair and unbiased.

Human oversight

- Automated decision-making is not used in the investigation of suspected illegal holiday rentals.
- The algorithm assists the employee in identifying the most probable cases of illegal holiday rental, which they can then prioritize for field investigation.
- A visualization of the algorithm's risk assessment is provided to the employee to help them decide whether to follow its recommendation or not.
- The supervisor and project enforcer are responsible for determining if there is actually a case of illegal housing.
- The algorithm has a significant influence on the planner, but it does not make independent decisions. Employees receive training to recognize the opportunities and risks of using algorithms.

More information

- City of Amsterdam private vacation rental service (in Dutch): https://www.amsterdam.nl/wonen-leefomgeving/wonen/vakantieverhuur/
- This system's Algorithm register entry: https://algoritmeregister.amsterdam.nl/en/illegal-holiday-rental-housing-risk/

Appendix

Amsterdam tourism policy

- The city of Amsterdam considers tourism as an important source of income and employment, but also acknowledges its negative impact on the livability of the city.
- The city is working on a sustainable tourism policy to ensure that tourism contributes to the quality of life in the city without causing overcrowding, nuisance, or damage to the environment.
- The measures being taken by the city include limiting the growth of hotels and holiday rentals in busy areas, promoting lesser-known neighborhoods and attractions, and stimulating sustainable modes of transport.
- The city also aims to tackle the nuisance caused by large groups of tourists and monitor the impact of tourism on the city to adjust policy if necessary.
- More information about the city's tourism policy can be found here (in Dutch): https://www.amsterdam.nl/bestuur-organisatie/volg-beleid/toerisme/

Random forest regression

- Random forest regression is a type of machine learning algorithm used for regression tasks, which involve predicting a continuous numerical value.
- It works by constructing many decision trees and aggregating their predictions to make a final prediction.
- Each decision tree is trained on a random subset of the available features and a random subset of the training data.
- The randomness introduced in building the trees helps to reduce the risk of overfitting and improves the accuracy of the model.
- Random forest regression is a popular and powerful tool for predictive modeling in many fields, including finance, healthcare, and engineering.

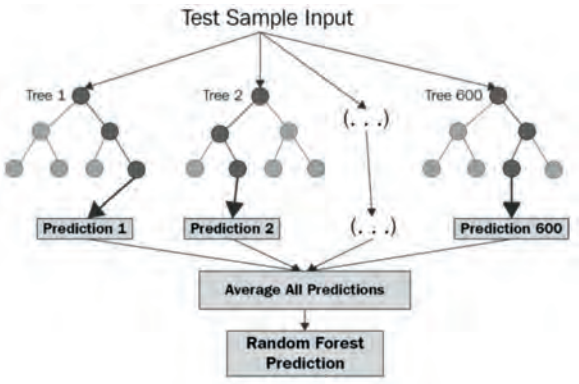


Figure E.2
A random forest regressor is constructed from multiple decision trees, the predictions of which are averaged. (Source: Keboola.)

SHAP

- SHapley Additive exPlanations (SHAP) is a method for explaining the output of machine learning models.
- It is based on game theory and uses the concept of Shapley values to assign importance to the input features that contribute to the output of the model.
- The method calculates the contribution of each feature by comparing the model's predictions with and without that feature.
- SHAP produces a set of explanations for each data point, which can help to interpret and understand the model's behavior.
- It is a flexible and model-agnostic method, meaning it can be applied to a wide range of machine learning models.
- More on SHAP: `https://shap.readthedocs.io`

Shapley values

- Shapley values are a concept from cooperative game theory.
- They measure the marginal contribution of each player in a coalition game.
- Shapley values have been adapted to machine learning as a way of assigning importance to input features in a model.
- In this context, the input features are treated as players in a game, and the output of the model is treated as the payoff.

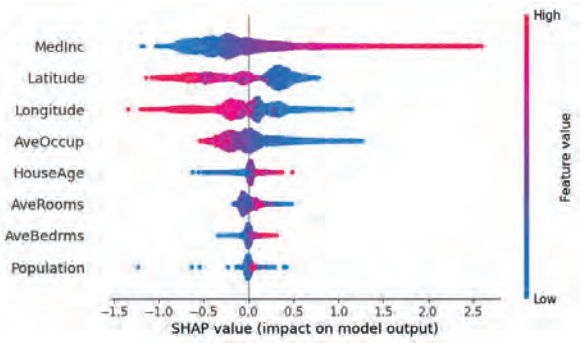


Figure E.3
Explaining feature impacts on model output of a housing price prediction random forest regressor using a SHAP beeswarm summary plot. (Source: Towards Data Science.)

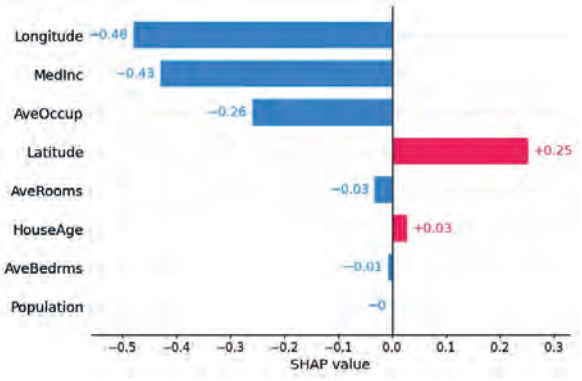


Figure E.4
Explaining feature impacts on a single housing price prediction by a random forest regressor using a SHAP bar plot. (Source: Towards Data Science.)

AI Fairness 360 toolkit

- The AI Fairness 360 (AIF360) is an open-source toolkit developed by IBM to help detect and mitigate bias in machine learning models.
- The toolkit includes a comprehensive set of metrics, algorithms, and tutorials that can be used to analyze and mitigate various forms of bias.
- The AIF360 toolkit can be used across various industries and domains to promote fair and trustworthy AI systems.

- The toolkit includes various components such as data preprocessing, bias detection, bias mitigation, and bias visualization.
- The AIF360 toolkit has been used in various real-world applications, including hiring and lending decisions, to ensure that AI systems are fair and unbiased.
- More on AIF360: `https://aif360.mybluemix.net`

Appendix F

Envisioning Contestability Loops: Focus Group Guide

Notes

- Make sure all results of the workshop are visible to all participants during the focus group.
- A focus group discussion can be relatively unstructured. But be sure to cover the points below in some depth.
- The discussion duration should be around 30–60 minutes.

Set-up

- Recap what we just did (the concept design workshop), and why (to apply the visual explanation to a real-world case).
- Introduce the focus group's purpose: to explore the group's experiences with the visual explanation.
- Go over the ground rules:
  - Put your phones on silent
  - If you need to leave (to take a call, go to the bathroom, etc.) do so quietly
  - Talk to each other, not just the moderator
  - There are no right or wrong answers
  - Feel free to disagree, but in a respectful manner
  - Try not to talk over each other, because this makes it harder to transcribe the recording later
  - The moderator can always interrupt to remind people of the rules
- Final questions before we begin?

☐ START RECORDING

Starting questions

- Do you use visual explanations in your practice (as a process tool)? Can you give any examples?
- What would you say makes a good visual explanation, or other design tool for inspiration, in general?
- How do you feel about the results of your concept design work just now in general? Are you satisfied? Why/why not?

The concept designs

- How do the design concepts compare to the properties of contestable AI systems described by the framework?
  - Interactive controls?
  - Intervention requests (incl. explanations)?
  - Tools for scrutiny?
  - Monitoring?
  - Participatory policy-making and systems development?
- Do the concept designs propose features that are not described in the framework and/or the visual explanation?

Working with the visual explanation

(Note: The questions below mirror the evaluation criteria from the form that the participants have completed individually before this focus group.)

- Did the visual explanation support your task? Why/why not?
- Was using the visual explanation easy to learn? Was it quick to use? Why/why not?
- Did the visual explanation contain all necessary information? If not, what was missing?
- Was the visual explanation adaptable to your process? Why/why not?
- Did the visual explanation inspire new ideas? Did it have the right level of abstraction? Why/why not?
- Was the visual explanation pleasant to use? Do you consider it attractive to look at? Why/why not?
- Are there any other strengths or weaknesses of the visual explanation that we did not cover?

Closing

- Why did you choose to join the workshop?
- What was it like to participate?
- Any other questions or comments?

Appendix G
Envisioning Contestability Loops: Concept Design Summaries

These summaries were auto-generated using ChatGPT (May 24 Version for workshops 1–3, and July 20 version for workshops 4–5)[1] with the following prompt:

> "Can you summarize the following description of a design concept for a contestable algorithmic system in 150 words or less, using plain and straightforward language?" [Raw transcript of design concept verbal description pasted on subsequent line.]

We then checked the summaries against the original transcripts and lightly edited them for correctness.

1. https://help.openai.com/en/articles/6825453-chatgpt-release-notes

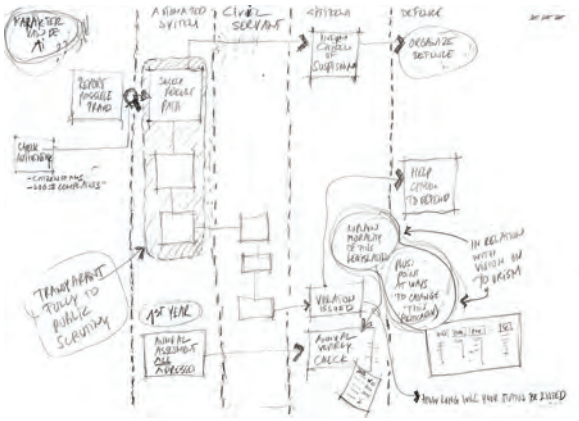


Figure G.1
Concept design 1 from workshop 1 (C1.1). Suggests implementing a transparent and fair approach to assess the behavior of citizens in Amsterdam. Initially, everyone starts with a clean status, and annual assessments are conducted for all citizens, regardless of whether they have been reported or not. The system aims to detect illegal renting practices, and individuals who report potential violations will be checked anonymously. The number of complaints filed by a person will be logged and considered in the selection process, as it indicates their behavior patterns. Immediate reports will be provided to individuals under investigation, allowing them to organize their defense early on. The algorithm's selection mechanism will be open to inspection by organizations like Bellingcat and Follow the Money. The annual assessment will provide individuals with information on their fraud listing status and an explanation of why the city of Amsterdam has chosen this approach. Additionally, the system will offer guidance on how individuals can challenge the rules and seek support from political parties if they disagree with the city's vision on tourism. The AI character, representing the system, should be carefully designed to ensure it portrays a fair and informative entity rather than an intimidating presence. (Summarized from 771 words.)

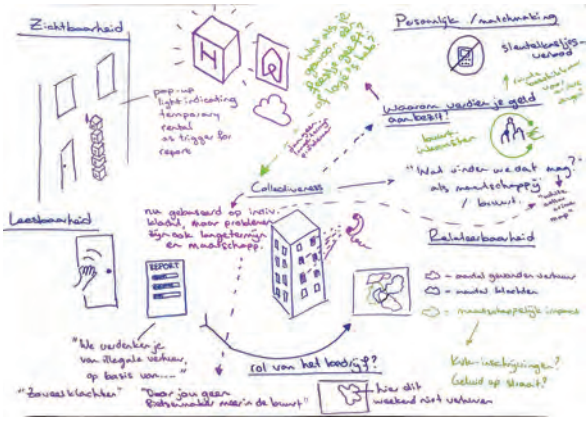


Figure G.2
Concept design 2 from workshop 1 (C1.2). Addresses issues related to complaints and the impact of platforms like Airbnb on neighborhoods. The idea involves creating a visible indicator outside homes, similar to hotel signs, to indicate if a property is being rented out. This would help people understand the situation before filing a complaint. Additionally, when authorities investigate a complaint, they would bring a report indicating the specific indicators that raised suspicion, such as previous incidents or neighborhood characteristics. The concept also emphasizes the need to consider collective effects beyond individual complaints, such as changes in neighborhoods due to short-term rentals. It proposes using data from sources like the Chamber of Commerce to understand broader societal trends rather than solely relying on individual statistics. The role of companies is also highlighted, suggesting they could discourage renting in already saturated neighborhoods. Personal matchmaking and the redistribution of profits from rentals are mentioned as additional considerations. Overall, the concept aims to balance individual concerns with collective impacts, enhance transparency, and encourage a more comprehensive approach to address issues related to short-term rentals. (Summarized from 672 words.)

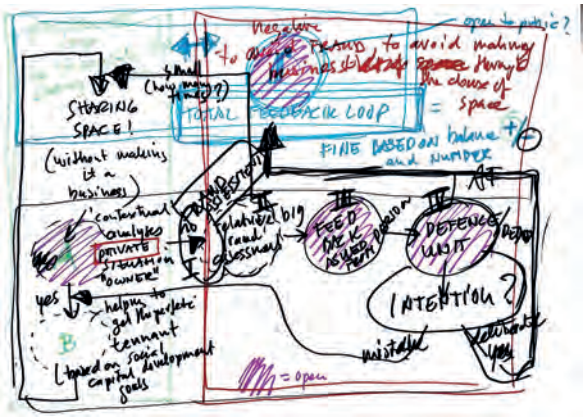


Figure G.3
Concept design 3 from workshop 1 (C1.3). Proposes a system for sharing valuable space in a positive and fair way. Instead of focusing on fraud detection, the idea is to encourage individuals to share their empty space with others. A contextual analysis is performed to determine if a person meets the conditions for sharing their space. If they do, the system helps them find a suitable tenant, aiming to bring together diverse individuals who wouldn't have otherwise met. The feedback loop suggests that any financial gains from this sharing could be shared among participants in some manner. To prevent misuse of the sharing space as a business opportunity, a fraud assessment is conducted. However, it's important to assess the severity of the offense, differentiating between minor and major infractions. Small mistakes allow individuals to restart the process, whereas deliberate or significant offenses result in immediate action. Rather than imposing immediate fines, the system considers an individual's past behavior and weighs their positive contributions against negative actions. This approach values and rewards those who consistently contribute positively over time. The concept also suggests the possibility of redistributing capital as a form of justice within the system. (Summarized from 691 words.)

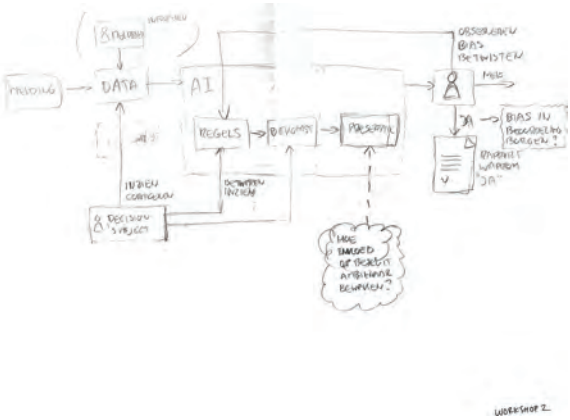


Figure G.4
Concept design from workshop 2 (C2). Begins with notifications, followed by data collection based on a specified description. The report is compiled from various sources and fed into an AI. The goal is to allow the "decision subject," such as a landlord or an individual affected by the decision, to have influence and oversight over the data collected about them. The system should incorporate rules governing the functioning of the AI and factors influencing its decision-making. Transparency and the ability to dispute the outcomes are important. The proposed outcome involves providing the decision subject with a visual representation of the official's perspective and the factors leading to the decision. However, the challenge lies in ensuring that the official's presentation remains unbiased, as the design itself can influence decisions. Additionally, the system should address biases that may arise from the civil servant's assessment. This interpretation aspect requires special attention to ensure its integrity. Regular notifications and visibility into AI outcomes allow individuals to contest the rules and processing methods. A solution is needed to address biases at both the AI and interpretation levels. (Summarized from 360 words.)

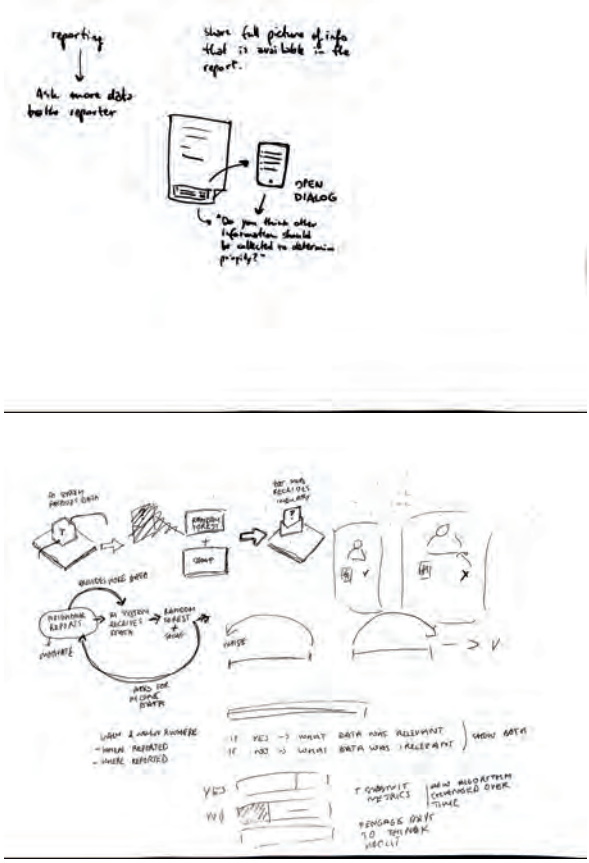


Figure G.5
Concept design from workshop 3 (C3). Focuses on transparency, dialogue, and feedback. The system aims to provide a full picture of information, including metrics and sources, to understand why a report is generated and how decisions are made. It encourages dialogue and feedback from both decision subjects and users to improve the system and make it more transparent. The concept involves involving developers and the public in system development and sharing success rates to engage them. It also addresses the communication and impact on individuals being investigated, emphasizing a human approach and minimizing negative effects. The design includes monitoring and collecting feedback, considering both the human and technical aspects. It also highlights the importance of results and ethical discussions while improving the system's fairness and effectiveness. Additionally, the concept suggests involving law enforcement for valuable insights and patterns, and exploring ways to account for errors and improve accuracy. The aim is to create an open, collaborative system that continuously improves with public input and helps achieve desired societal goals. (Summarized from 3090 words.)

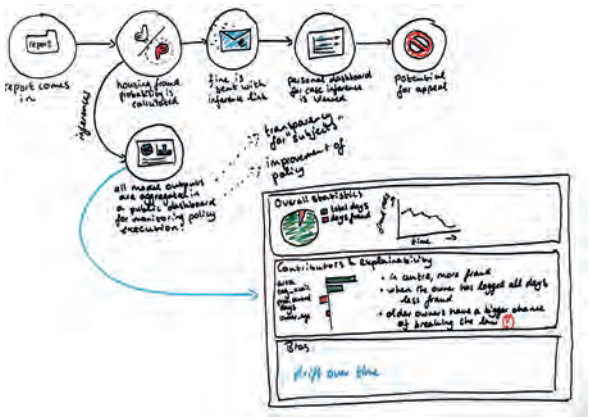


Figure G.6
Concept design 1 from workshop 4 (C4.1). Aims to address fraud cases and provide transparency and accountability. When a report indicates potential fraud, a fine is issued, and the person can access an "inference dashboard" showing the factors influencing the decision. The key addition is a "monitoring dashboard" that operates on an aggregated scale, visible to both policymakers and the public. This dashboard has three parts: (1) Overall statistics: Showing the proportion of normal days versus fraud cases over time, giving context and aiding policy adjustments. (2) Contributors: Highlighting features and their impact on the system (e.g., location or age) with explanatory statements for better policymaking. (3) Bias overview: Monitoring model drift and bias evolution over time to identify potential issues. By implementing this system, it becomes possible to steer policy decisions based on data, promote fairness, and build public trust in algorithmic processes. (Summarized from 388 words.)

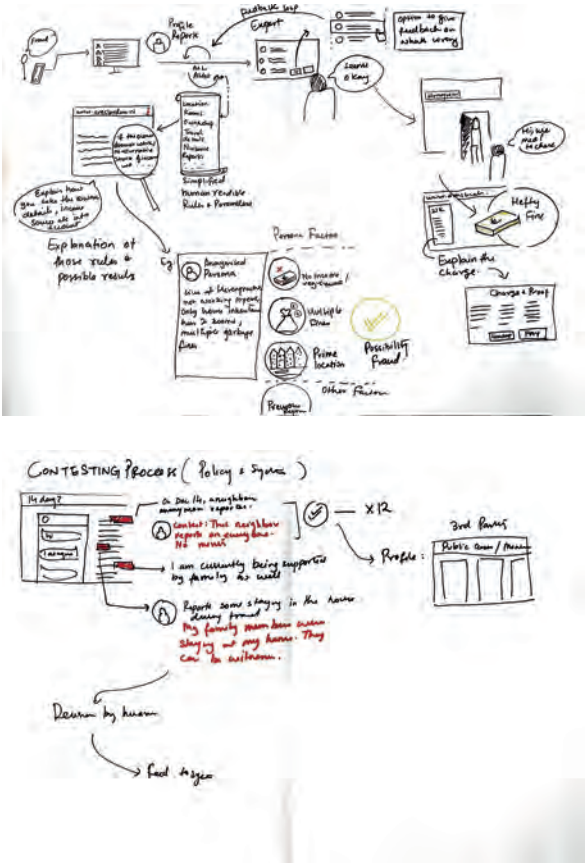


Figure G.7
Concept design 2 from workshop 4 (C4.2). Focuses on creating an explainable and transparent process for handling fraud reports. The system aims to allow citizens and experts to understand how the algorithms work and how decisions are made. To achieve this, the system would route algorithm parameters and results to a website where they can be presented in a more human-readable manner, using anonymized personas to illustrate examples. The process involves human intervention, where experts can review algorithm-generated results and provide qualitative feedback, which is then recorded in the system. If a charge is levied based on the algorithm's assessment, the system ensures that the charge is clearly explained, and the individual has the option to contest it before paying. The contesting process is designed to allow citizens to challenge decisions with relevant evidence. However, some challenges remain, such as determining time limits for contestation and addressing potential biases in the system. Implementing the system requires considering policy implications and finding ways to aggregate and present bias-related data to address recurring patterns. Overall, the design concept aims to empower citizens by giving them agency in influencing decisions made by the algorithmic system while ensuring transparency, fairness, and accountability throughout the process. (Summarized from 1197 words.)

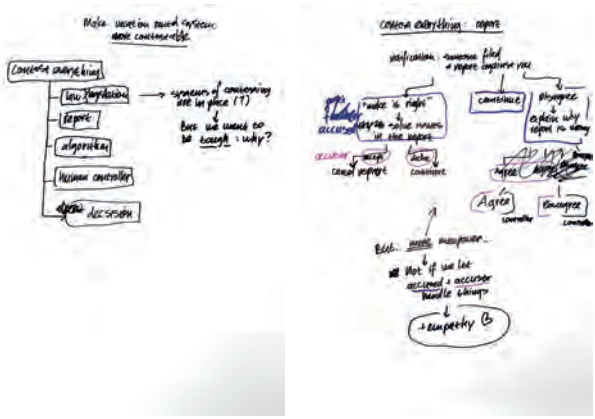


Figure G.8
Concept design 3 from workshop 4 (C4.3). Involves contesting various aspects of the process. It begins with legislation, suggesting the need for more empathy and understanding rather than just imposing fines. The second step is contesting the report, where the accused is notified and given options to make things right before the process starts. The accuser can also receive feedback and decide whether to proceed or not. Instead of relying solely on a human controller, the idea is to let the accuser and accused work things out together, fostering empathy and understanding during the process. The final aspect is contesting the algorithm, allowing individuals to challenge the analysis provided by the algorithm along with the report. This approach aims to improve the system's fairness and effectiveness while promoting collaboration and empathy between parties involved. (Summarized from 643 words.)

Figure G.9
Concept design 1 from workshop 5 (C5.1). Involves identifying vulnerable individuals related to Airbnb rentals and using a step-by-step process to handle potential issues. The primary target is "Joke," an Airbnb host who unintentionally commits fraud by forgetting her registration number. The system triggers an alert and provides instructions for her to rectify the situation. If she fails to respond or disputes the decision, a handhaver (enforcement officer) intervenes. During this process, there are opportunities for her to provide feedback and challenge the algorithm's decisions. It's essential to give enough time between steps to accommodate adjustments and feedback. The system aims for transparent decision-making and swift resolution while considering the user's circumstances and ensuring fairness. (Summarized from 949 words.)

Figure G.10
Concept design 2 from workshop 5 (C5.2). Involves a circular process. It begins with collecting reports of potential issues or disturbances. The system aims to make this information transparent to the affected individuals, allowing them to verify its accuracy. Using the collected data about the property and people involved, the model then assesses the likelihood of fraud. When a staff member reviews the case, they have access to the prediction's certainty level, aided by explanations from SHAP. The system facilitates communication with the affected person without immediately accusing them of fraud, allowing for validation and potential corrections. The decision-making process also considers human judgment and explores whether policies need adjustment. Two key stages are determining whether to initiate an investigation and deciding whether to impose a penalty, with a focus on ensuring proportionate consequences for intentional versus unintentional errors. The system aims to provide insight into the rules while minimizing administrative burdens for individuals and allowing intervention when needed. (Summarized from 863 words.)


P