Envisioning Contestability Loops

Evaluating the Agonistic Arena as a Generative Metaphor for Public AI

Alfrink, Kars; Keller, Ianus; Yurrita Semperena, Mireia; Bulygin, Denis; Kortuem, Gerd; Doorn, Neelke

DOI
10.1016/j.sheji.2024.03.003

Publication date
2024

Document Version
Final published version

Published in
She Ji

she ji
[ʃə: ʤi:]

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

Kars Alfrink
Ianus Keller
Mireia Yurrita Semperena
Denis Bulygin
Gerd Kortuem
Neelke Doorn

Keywords
artificial intelligence
contestability
generative metaphor
interaction design
public administration
visual explanations

Received
October 24, 2023
Accepted
March 25, 2024

KARS ALFRINK
Department of Sustainable Design Engineering, Delft University of Technology, The Netherlands
c.p.alfrink@tudelft.nl

IANUS KELLER
Department of Human Centered Design, Delft University of Technology, The Netherlands
a.i.keller@tudelft.nl

MIREIA YURRITA SEMPERENA
Department of Sustainable Design Engineering, Delft University of Technology, The Netherlands
m.yurritasemperena@tudelft.nl

DENIS BULYGIN
Department of Sustainable Design Engineering, Delft University of Technology, The Netherlands
d.bulygin@tudelft.nl

Abstract
Public sector organizations increasingly use artificial intelligence to augment, support, and automate decision-making. However, such public AI can potentially infringe on citizens' right to autonomy. Contestability is a system quality that protects against this by ensuring systems are open and responsive to disputes throughout their life cycle. While a growing body of work is investigating contestable AI by design, 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, inspired by the political theory of agonistic pluralism. Combining this metaphor and current contestable AI guidelines, we develop 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 outline the mechanisms for contestability derived from 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.

http://www.sciencedirect.com/journal/she-ji-the-journal-of-design-economics-and-innovation
https://doi.org/10.1016/j.sheji.2024.03.003

Introduction

Algorithmic decision-making in the public sector can undermine autonomy—people's effective capacity for self-governance.[1] If we are to guard against such undermining, public artificial intelligence (AI) systems should be *contestable*: open and responsive to disputes throughout their lifecycle, establishing dialogical relationships between decision subjects (people who are significantly impacted by human-AI system actions) and system operators.[2]

Contestable AI is an emerging field of research within human-centered AI.[3] However, as with other aspects of responsible AI, much of the debate related to contestability has been focused on principles rather than practices.[4] For practitioners to use the findings from the contestable AI field, they need to be translated and adapted to specific contexts[5] and presented in ways to which they can easily relate.[6] One such form is visual explanations—infographics that represent dynamic processes.[7] Furthermore, design knowledge should be generative—allowing for a range of specific solutions without entirely prescribing their form.[8] We can achieve such conceptual richness by articulating a generative metaphor,[9] which is an idea that allows designers to think of a problem in terms of something else, leading to particular diagnoses and accompanying prescriptions.

In this article, we hypothesize a generative metaphor for contestable AI in the public sector: the Agonistic Arena. We developed this metaphor based on the concept of agonistic pluralism,[10] a political philosophy that underpins much contestable AI research. Our main aim was to evaluate the efficacy of the Agonistic Arena metaphor as a generative metaphor for designing more contestable public AI. In support of this aim, we created an infographic of contestable AI that supports practitioners during the concept design of public AI, titled "Contestability Loops for Public AI." The infographic is built on previous work, translating contestable AI into more practical guidance.[11] It was also deliberately designed to convey the Agonistic Arena metaphor. We qualitatively evaluated this infographic with practicing designers in a series of workshops. In these workshops, we asked participants 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.[12] Our ontological and epistemological commitments are critical realist[13] and contextualist.[14] We used creative design to produce an artifact and use it as the research instrument to generate the data. The data was analyzed using interpretative techniques.

The contributions to the field of design from this study are: the construction of the Agonistic Arena based on political theory, a generative metaphor that animates the contestable AI field; an infographic that further concretizes and explicates contestable AI knowledge for the audience of design practitioners active in the public AI space; an evaluation of whether and to what extent practicing designers produce more contestable concept designs of public AI when using the Arena metaphor and the Contestability Loops infographic; and an account of several competing metaphors that may be at play in public AI discourse—the *Black Box* and the *Sovereign*.

1 Alan Rubel, Clinton Castro, and Adam K. Pham, *Algorithms and Autonomy: The Ethics of Automated Decision Systems* (Cambridge: Cambridge University Press, 2021), https://doi.org/10.1017/9781108895057; Carina Prunkl, "Human Autonomy in the Age of Artificial Intelligence," *Nature Machine Intelligence* 4 (February 2022): 99–101, https://doi.org/10.1038/s42256-022-00449-9; Sábëlo Mhlambi and Simona Tiribelli, "Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms," *Topoi* 42 (February 2023): 867–80, https://doi.org/10.1007/s11245-022-09874-2.

2 Kars Alfrink et al., "Contestable AI by Design: Towards a Framework," *Minds and Machines* 33 (August 2022): 613–39, https://doi.org/10.1007/s11023-022-09611-z.

3 Tara Capel and Margot Brereton, "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* (New York: ACM, 2023), paper no. 359, https://doi.org/10.1145/3544548.3580959.

4 Jessica Morley et al., "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics* 26 (December 2019): 2141–68, https://doi.org/10.1007/s11948-019-00165-5.

5 Colin M. Gray and Yubo Kou, "UX Practitioners' Engagement with Intermediate-Level Knowledge," in *DIS '17 Companion: Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems* (New York: ACM, 2017), 13–17, https://doi.org/10.1145/3064857.3079110.

6 Jennifer Williams, Dena Fam, and Abby Mellick Lopes, "Creating Knowledge:

GERD KORTUEM
Department of Sustainable Design Engineering, Delft University of Technology, The Netherlands
g.w.kortuem@tudelft.nl

NEELKE DOORN
Department of Values, Technology and Innovation, Delft University of Technology, The Netherlands
n.doorn@tudelft.nl

This article is structured as follows: First, we provide background on public AI, contestable AI, generative metaphor, agonistic pluralism, and the Agonistic Arena metaphor. Second, we describe our method, including infographic design, workshop focus groups, and reflexive thematic analysis. Third, 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.

Background

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.[15]

Research on the use of AI in the public sector (public AI) is growing.[16] While some use AI,[17] the terms algorithm or algorithmic system are more prevalent.[18] Such systems inform or automate government decision-making.[19] Key application areas of public AI are child protection, public housing, health, social protection, security, and taxation.[20] The primary concerns in public AI include transparency,[21] data collection politics,[22] and impact on public sector work.[23]

A related field is urban AI,[24] which delves into the role of AI in the built environment. Here, the emphasis is on mobility solutions such as electric vehicle charging, autonomous vehicles, and parking systems.[25] This research examines the influence of AI on urban experiences, intertwining AI ethics with urban design ethics.[26] The focus on spatial justice[27] is unique to urban AI, complementing procedural and distributive justice discussions.

One of the issues relevant to public AI is that of autonomy,[28] which the emerging field of contestable AI seeks to address.

Contestable AI

Research on contestable AI has been expanding, highlighting its significance in safeguarding against flawed and unjust automated decision-making. It emphasizes human involvement and the fostering of adversarial discussions between decision subjects and system operators.[29]

Contestability can be viewed as humans questioning machine predictions, allowing human intervention to rectify potential machine errors.[30] It can be described as a blend of human and machine decision-making, emphasizing its role in procedural justice and enhancing perceived fairness.[31] The practice of "contestability by design" stresses human intervention retrospectively and in the AI development processes.[32] Contestability transcends mere human intervention, demanding a dialectical interaction between decision subjects and human controllers.[33] The legitimacy of a system is compromised without contestability, which in turn demands both explanations and justifications.[34] Implementing contestability features in practice requires thoughtful consideration of needs, values, and context.[35]

Visual Communication Design Research in Transdisciplinary Projects," in *Transdisciplinary Research and Practice for Sustainability Outcome*, ed. Dena Fam et al. (London: Routledge, 2016), chapter 11, https://doi.org/10.4324/9781315652184.

7 Edward R. Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative* (Cheshire, CT: Graphics Press, 1997).

8 Kristina Höök and Jonas Löwgren, "Strong Concepts: Intermediate-Level Knowledge in Interaction Design Research," *ACM Transactions on Computer-Human Interaction* 19, no. 3 (2012): 1–18, https://doi.org/10.1145/2362364.2362371.

9 Donald A. Schön, "Generative Metaphor: A Perspective on Problem-Setting in Social Policy," in *Metaphor and Thought*, 2nd ed., ed. Andrew Ortony (Cambridge: Cambridge University Press, 1993), 137–63, https://doi.org/10.1017/CBO9781139173865.011.

10 Chantal Mouffe, *The Return of the Political* (London: Verso, 1993).

11 Alfrink et al., "Contestable AI by Design"; Kars Alfrink et al., "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* (New York: ACM, 2023), article no. 8, https://doi.org/10.1145/3544548.3580984.

12 Ilpo Koskinen, Thomas Binder, and Johan Redström, "Lab, Field, Gallery, and Beyond," *Artifact* 2, no. 1 (2008): 46–57, https://doi.org/10.1080/17493460802303333; Ilpo Koskinen et al., *Design Research through Practice: From the Lab, Field, and Showroom* (Boston: Morgan Kaufmann, 2012), https://doi.org/10.1016/B978-0-12-385502-2.00013-4.

13 Christopher Frauenberger, "Critical Realist HCI," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (New York: ACM, 2016), 341–51, https://doi.org/10.1145/2851581.2892569; Philip S. Gorski, "'What Is Critical Realism? And Why Should You Care?'" *Contemporary Sociology: A Journal of Reviews* 42, no. 5 (2013): 658–70, https://doi.org/10.1177/0094306113499533.

14 Anna Madill, Abbie Jordan, and Caroline Shirley, "Objectivity and Reliability in Qualitative Analysis: Realist, Contextualist and Radical Constructionist Epistemologies," *British Journal of Psychology* 91, no. 1 (2000): 1–20, https://doi.org/10.1348/000712600161646; Stuart D. Green, Chung-Chin Kao, and Graeme D. Larsen, "Contextualist Research: Iterating between Methods While Following an Empirically Grounded Approach," *Journal of Construction Engineering*

and Management 136, no. 1 (2010): 117–26, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000027.

15 Sem Nouws, Marijn Janssen, and Roel Dobbe, "Dismantling Digital Cages: Examining Design Practices for Public Algorithmic Systems," in *Electronic Government: EGOV 2022, Lecture Notes in Computer Science*, vol. 13391, ed. Marijn Janssen et al. (Cham: Springer-Verlag, 2022), 307–22, https://doi.org/10.1007/978-3-031-15086-9_20; Lucy Suchman, "Corporate Accountability," *Robot Futures*, June 10, 2018, https://robotfutures.wordpress.com/2018/06/10/corporate-accountability/.

16 Anna Brown et al., "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* (New York: ACM, 2019), paper no. 41, https://doi.org/10.1145/3290605.3300271; Karolina Drobotowicz, Marjo Kauppinen, and Sari Kujala, "Trustworthy AI Services in the Public Sector: What Are Citizens Saying about It?," in *Requirements Engineering: Foundation for Software Quality*, ed. Fabiano Dalpiaz and Paola Spoletini (Cham: Springer, 2021), 99–115, https://doi.org/10.1007/978-3-030-73128-1_7; Karl de Fine Licht and Jenny de Fine Licht, "Artificial Intelligence, Transparency, and Public Decision-Making: Why Explanations Are Key When Trying to Produce Perceived Legitimacy," *AI & SOCIETY* 35 (December 2020): 917–26, https://doi.org/10.1007/s00146-020-00960-w; Samar Fatima et al., "Public AI Canvas for AI-enabled Public Value: A Design Science Approach," *Government Information Quarterly* 39, no. 4 (2022): 101722, https://doi.org/10.1016/j.giq.2022.101722; Asbjørn Ammitzbøll Flügge, "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* (New York: ACM, 2021), 253–55, https://doi.org/10.1145/3462204.3481787; Vidushi Marda and Shivangi Narayan, "Data in New Delhi's Predictive Policing System," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York: ACM, 2020), 317–24, https://doi.org/10.1145/3351095.3372865; Kathleen H. Pine and Max Liboiron, "The Politics of Measurement and Action," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing*

In this study, we conceptualize contestable AI as systems that are open to human intervention throughout their lifecycle, emphasizing a dialogical relationship with decision subjects. We can leverage these contestations for continuous system improvement.[36] A proposed design framework lists elements contributing to contestability, incorporating system features and development practices tied to stakeholders and AI system lifecycle stages.[37] Subsequent work emphasizes the relevance of participatory policy-making approaches and the need to monitor contestations for systemic flaws.[38]

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 prescriptions.

Generative Metaphor

Donald Schön defines generative metaphor as a lens influencing our perception and understanding of the world.[39] 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.[40] Recognizing society's implicit generative metaphors enhances our understanding. Not all metaphors are generative; only those offering new insights qualify. Schön also discusses frame restructuring as a means to reconcile conflicting perspectives.[41]

Related yet distinct is George Lakoff's theory of conceptual metaphor.[42] 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.[43] In essence, metaphorical thought is unavoidable.

In human-computer interaction (HCI) and design research, researchers have used generative metaphors to analyze discourses in computing,[44] user perception of voice interfaces,[45] and to challenge HCI research assumptions.[46] Attempts to formalize the methodical use of metaphor include Method Cards,[47] which helpfully categorize them as weak or strong. Functional prototypes in various domains use metaphor for design, including AI.[48] Metaphor and narrative synergistically enhance moral imagination, offering a dynamic approach to the value-sensitive design of AI systems.[49] Metaphorical thinking can foster a more nuanced understanding of artificial intelligence.[50] Designers cannot escape the use of metaphor, and it is best that they do this consciously.[51]

Generative metaphor shows that design issues can be interpreted in multiple ways. Each interpretation suggests particular underlying challenges. Therefore, how we metaphorically frame AI problems matters. Understanding a driving metaphor underpins the effective use of prescriptions of a more tactical nature. The following section describes the metaphor we feel best guides contestable AI thinking.

Systems (New York: ACM, 2015), 3147–56, https://doi.org/10.1145/2702123.2702298; Devansh Saxena and Shion Guha, "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* (New York: ACM, 2020), 383–88, https://doi.org/10.1145/3406865.3418331; Devansh Saxena et al., "A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare," *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (2021): article no. 348, https://doi.org/10.1145/3476089; Michael Veale, Max van Kleek, and Reuben Binns, "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* (New York: ACM, 2018), paper no. 440, https://doi.org/10.1145/3173574.3174014.

17 Drobotowicz et al., "Trustworthy AI Services"; Fine Licht and Fine Licht, "Artificial Intelligence, Transparency, and Public Decision-Making"; Fatima et al., "Public AI Canvas."

18 Brown et al., "Toward Algorithmic Accountability"; Saxena et al., "Framework of High-Stakes Algorithmic Decision-Making"; Flügge, "Perspectives from Practice"; Saxena and Guha, "Conducting Participatory Design"; Veale et al., "Fairness and Accountability Design Needs"; Pine and Liboiron, "Politics of Measurement and Action."

19 Brown et al., "Toward Algorithmic Accountability"; Drobotowicz et al., "Trustworthy AI Services"; Fatima et al., "Public AI Canvas"; Saxena and Guha, "Conducting Participatory Design"; Saxena et al., "Framework of High-Stakes Algorithmic Decision-Making."

20 Brown et al., "Toward Algorithmic Accountability"; Drobotowicz et al., "Trustworthy AI Services"; Marda and Narayan, "Data in New Delhi's Predictive Policing System."

21 Brown et al., "Toward Algorithmic Accountability"; Drobotowicz et al., "Trustworthy AI Services"; Fine Licht and Fine Licht, "Artificial Intelligence, Transparency, and Public Decision-Making."

22 Marda and Narayan, "Data in New Delhi's Predictive Policing System"; Pine and Liboiron, "Politics of Measurement and Action."

23 Flügge, "Perspectives from Practice"; Saxena and Guha, "Conducting Participatory Design"; Saxena et al., "Framework of High-Stakes Algorithmic Decision-Making"; Veale et al., "Fairness and Accountability Design Needs."

24 Aale Luusua and Johanna Ylipulli, "Urban AI: Formulating an Agenda for the Interdisciplinary Research of Artificial Intelligence in Cities," in *DIS' 20 Companion: Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (New York: ACM, 2020), 373–76, https://doi.org/10.1145/3393914.3395905; Aale Luusua et al., "Urban AI: Understanding the Emerging Role of Artificial Intelligence in Smart Cities," *AI & SOCIETY* 38 (June 2023): 1039–44, https://doi.org/10.1007/s00146-022-01537-5; Federico Cugurullo, "Urban Artificial Intelligence: From Automation to Autonomy in the Smart City," *Frontiers in Sustainable Cities* 2 (July 2020): 1–14, https://doi.org/10.3389/frsc.2020.00038.

25 Kars Alfrink et al., "Tensions in Transparent Urban AI: Designing a Smart Electric Vehicle Charge Point," *AI & Society* 3 (June 2022): 1049–65, https://doi.org/10.1007/s00146-022-01436-9; Aale Luusua and Johanna Ylipulli, "Artificial Intelligence and Risk in Design," in *DIS '20: Proceedings of the 2020 ACM Designing Interactive Systems Conference* (New York: ACM, 2020), 1235–44, https://doi.org/10.1145/3357236.3395491; Nitin Sawhney, "Contestations in Urban Mobility: Rights, Risks, and Responsibilities for Urban AI," *AI & SOCIETY* 38 (June 2023): 1083–98, https://doi.org/10.1007/s00146-022-01502-2.

26 Luusua and Ylipulli, "Urban AI."

27 Henri Lefebvre, "Le Droit à La Ville," *L'Homme Et La Société* 6, no. 1 (1967): 29–35, available at https://www.persee.fr/doc/homso_0018-4306_1967_num_6_1_1063; Edward Soja, "The City and Spatial Justice," Justice Spatiale/Spatial Justice 1, no. 1 (2009): 1–5, https://www.jssj.org/article/la-ville-et-la-justice-spatiale/?lang=en; Joe Shaw and Mark Graham, "An Informational Right to the City? Code, Content, Control, and the Urbanization of Information," *Antipode* 49, no. 4 (2017): 907–27, https://doi.org/10.1111/anti.12312; David Harvey, "The Right to the City," *International Journal of Urban and Regional Research* 27, no. 4 (2003): 939–41, https://doi.org/10.1111/j.0309-1317.2003.00492.x.

28 Rubel et al., *Algorithms and Autonomy*; Prunkl, "Human Autonomy"; Mhlambi and Tiribelli, "Decolonizing AI Ethics."

29 Capel and Brereton, "What Is Human-Centered"; Tad Hirsch et al., "Designing Contestability: Interaction Design, Machine Learning, and Mental Health," in *DIS '17: Proceedings of the 2017 Conference on Designing Interactive Systems* (New York: ACM, 2017), 95–99, https://doi.org/10.1145/3064663.3064703; Marco Almada, "Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems," in *ICAIL '19: Proceedings of the 17th International Conference on Artificial Intelligence and Law* (New York: ACM, 2019), 2–11, https://doi.org/10.1145/3322640.3326699; Kristen Vaccaro et al., "Contestability in Algorithmic Systems," in *CSCW '19 Companion: Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing* (New York: ACM, 2019), 523–27, https://doi.org/10.1145/3311957.3359435; Claudio Sarra, "Put Dialectics into the Machine: Protection against Automatic-Decision-Making through a Deeper Understanding of Contestability by Design," *Global Jurist* 20, no. 3 (2020): 20200003, https://doi.org/10.1515/gj-2020-0003; Clément Henin and Daniel Le Métayer, "Beyond Explainability: Justifiability and Contestability of Algorithmic Decision Systems," *AI & SOCIETY* 37 (December 2022): 1397–1410, https://doi.org/10.1007/s00146-021-01251-8; Alfrink et al., "Contestable AI by Design."

30 Hirsch et al., "Designing Contestability"; Himanshu Verma et al., "Rethinking the Role of AI with Physicians in Oncology: Revealing Perspectives from Clinical and Research Workflows," in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2023), article no. 17, https://doi.org/10.1145/3544548.3581506.

31 Vaccaro et al., "Contestability in Algorithmic Systems"; Mireia Yurrita et al., "Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability," in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2023), article no. 134, https://doi.org/10.1145/3544548.3581161; Henrietta Lyons, Tim Miller, and Eduardo Velloso, "Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York: ACM, 2023), 764–74, https://doi.org/10.1145/3593013.3594041.

32 Almada, "Human Intervention."

33 Sarra, "Put Dialectics into the Machine."

34 Henin and Le Métayer, "Beyond Explainability."

35 Henrietta Lyons, Eduardo Velloso, and Tim Miller, "Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions," *Proceedings of the ACM on*

Human-Computer Interaction 5, no. CSCW1 (2021): article no. 106, https://doi.org/10.1145/3449180.

36 Alfrink et al., "Contestable AI by Design."

37 Ibid.

38 Alfrink et al., "Contestable Camera Cars."

39 Schön, "Generative Metaphor," 138.

40 Ibid., 139.

41 Ibid., 150–161.

42 George Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* (Chicago: University of Chicago Press, 1987).

43 Ibid., 3–5.

44 Anne Hamilton, "Metaphor in Theory and Practice: The Influence of Metaphors on Expectations," *ACM Journal of Computer Documentation* 24, no. 4 (2000): 237–53, https://doi.org/10.1145/353927.353935; Maria Lindh, "As a Utility — Metaphors of Information Technologies," *Human IT: Journal for Information Technology Studies as a Human Science* 13, no. 2 (2016): 47–80, https://humanit.hb.se/article/view/418.

45 Smit Desai and Michael Twidale, "Metaphors in Voice User Interfaces: A Slippery Fish," *ACM Transactions on Computer-Human Interaction* 30, no. 6 (2023): article no. 89, https://doi.org/10.1145/3609326.

46 Jordan Beck and Hamid R. Ekbia, "The Theory-Practice Gap as Generative Metaphor," in *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2018), paper no. 620, https://doi.org/10.1145/3173574.3174194.

47 Nick Logler, Daisy Yoo, and Batya Friedman, "Metaphor Cards: A How-to-Guide for Making and Using a Generative Metaphorical Design Toolkit," in *DIS '18: Proceedings of the 2018 Designing Interactive Systems Conference* (New York: ACM, 2018), 1373–86, https://doi.org/10.1145/3196709.3196811.

48 Graham Dove and Anne-Laure Fayard, "Monsters, Metaphors, and Machine Learning," in *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2020), 1–17, https://doi.org/10.1145/3313831.3376275; Dave Murray-Rust, Iohanna Nicenboim, and Dan Lockton, "Metaphors for Designers Working with AI," in *DRS Biennial Conference Series*, ed. Dan Lockton et al. (Bilbao, Spain: DRS, 2022), 1–20, https://doi.org/10.21606/drs.2022.667; Iohanna Nicenboim et al., "Conversation Starters: How Can We Misunderstand AI Better?," in *CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*

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," which takes inspiration from the ancient Greek ideal of democratic competitiveness.[52] This metaphor casts AI systems as a space in which conflict in various forms is embraced and celebrated as a productive force. Agonistic pluralism is the political philosophy underpinning this metaphor.

Agonistic pluralism, developed by Chantal Mouffe,[53] 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 challenging 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.[54]

In science and technology studies (STS), agonistic pluralism is employed to critique participation and inclusion approaches in responsible research and innovation(RRI). Jack Stilgoe, Richard Owen, and Phil Macnaghten discuss the limitations of inclusion in responsible research and innovation, suggesting that it often becomes an end in itself, shaped by those in power, and overlooks the diverse motivations of participants.[55] They advocate for more critical reflection on participation and its underlying norms. Jeroen van Bouwel and Michiel van Oudheusden argue for a differentiated approach to democratizing scientific governance; they point out that consensus in democracy often neglects conflict and non-consensual change, and they thus advocate for models like agonistic pluralism that embrace disagreement.[56] Audley Genus and Andy Stirling highlight the importance of inclusive, reflexive deliberation in responsible research and innovation, acknowledging the challenges posed by dogmatism and advocating for incrementalism.[57] Eugen Popa, Vincent Blok, and Renate Wesselink focus on the role of conflict in technology history, proposing agonism to manage conflict by valuing responsiveness and dialogue over consensus.[58] Similarly, Deborah Scott observes that challenges in public engagement in responsible research and innovation reflect criticisms of deliberative democracy and suggest an "agonistic" responsible research and innovation that examines power relations and views stakeholder stances as adversarial rather than equally valid.[59]

Researchers have applied agonistic pluralism in the context of interaction design, AI, machine learning (ML), and algorithmic decision-making.[60]

(New York: ACM, 2023), article no. 431, https://doi.org/10.1145/3544549.3583914; Jesse Josua Benjamin et al., "The Entoptic Field Camera as Metaphor-Driven Research-Through-Design with AI Technologies," in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2023), article no. 178, https://doi.org/10.1145/3544548.3581175.

49 Steven Umbrello, "Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design," *Science and Engineering Ethics* 26 (April 2020): 575–95, https://doi.org/10.1007/s11948-019-00104-4.

50 Filippo Fabrocini and Kostas Terzidis, "Re-framing AI: An AI Product Designer Perspective," *Techné: Research in Philosophy and Technology* 25, no. 3 (2021): 407–33, https://doi.org/10.5840/techne2021127151.

51 Paul Hekkert and Nazlı Cila, "Handle with Care! Why and How Designers Make Use of Product Metaphors," *Design Studies* 40 (September 2015): 196–217, https://doi.org/10.1016/j.destud.2015.06.007.

52 Jakub Filonik, "We Are the Champions: The Role of Agonistic Metaphor in the Political Discourse of Classical Greece," in *The* Agōn *in Classical Literature: Studies in Honour of Professor Chris Carey*, ed. Michael Edwards et al. (London: University of London Press, 2022), 155–62, https://doi.org/10.14296/wkue3508.

53 Mouffe, *Return of the Political*; Chantal Mouffe, *Agonistics: Thinking the World Politically* (London: Verso, 2013); Chantal Mouffe, "Deliberative Democracy or Agonistic Pluralism?," *Social Research* 66, no. 3 (1999): 745, https://www.jstor.org/stable/40971349; Chantal Mouffe, *The Democratic Paradox* (London: Verso, 2000); Chantal Mouffe, "Pluralism, Dissensus and Democratic Citizenship," in *Education and the Good Society*, ed. Fred Inglis (New York: Springer, 2004), 42–53; Chantal Mouffe, *On the Political* (London: Routledge, 2005); Chantal Mouffe, "Some Reflections on an Agonistic Approach to the Public," in *Making Things Public: Atmospheres of Democracy*, ed. Bruno Latour and Peter Weibel (Cambridge, MA: MIT Press, 2005), 804–7.

54 Carl DiSalvo, "Design, Democracy and Agonistic Pluralism," in *Design and Complexity — DRS International Conference 2010*, ed. David Durling et al. (Montreal, Canada: DRS, 2010), https://dl.designresearchsociety.org/drs-conference-papers/drs2010/researchpapers/31; Vivien Lowndes and Marie Paxton, "Can Agonism Be Institutionalised? Can

AI systems seen as objects of agonistic political design create spaces for confronting power relations.[61] Adversarial design methods can democratize technology development in line with agonistic ideals.[62] The agonistic lens helps us see that AI systems are also always part of contested spaces. When properly agonistic, algorithmic decision-making is always a provisional and temporary stabilization of power.[63] Agonistic AI system development would allow people to decide if, when, and how to integrate AI. Agonistic AI decision-making allows individuals to demand alternative ways of being computed or entirely reject being computed.[64] Agonistic AI demands broader forms of participation that acknowledge and allow for conflict and are sensitive to power relations and exclusions.[65] Agonism lets us see AI systems not only as a product or producer of politics but also as a space within which politics happens and to resist simplistic readings of AI's politics as fully liberatory or oppressive.[66] In contrast to AI safety approaches that rely on principles or technologies, AI development can be conceived of as machine politics, where agonistic deliberation should aim not merely to achieve AI safety but also to embody its goal.[67]

Conceptualizing the Generative Metaphor of the "Agonistic Arena"

Contestable AI embodies the generative metaphor of the Arena. This metaphor characterizes public AI as a space where interlocutors embrace conflict as productive. Seen through the lens of the Arena, public AI problems stem from a need for opportunities for adversarial interaction between stakeholders. This metaphorical framing suggests prescriptions to make more contentious and open to dispute the norms and procedures that shape 1) AI system design decisions on a global level and 2) human-AI system output decisions on a local level (i.e., 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.

Design and AI

Our empirical work centered on early-stage design activities focused on generating concept designs; this does not mean we hold a linear deterministic view of how design contributes to AI systems. Actually existing AI systems are designed and redesigned continuously by people whose job titles do not include the word designer, who do not consider themselves doing design, 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 and constantly changing in response to feedback from their environment.[68]

In this context, design is more akin to what John Seely Brown described as "thinkering" — experimenting, testing, and adjusting in a collaborative manner similar to the open-source approach.[69] Malcolm McCullough has described this as tuning — the incremental growth, change, and adaptation of configurations and settings based on the *feel* of the aggregate, something

Institutions Be Agonised? Prospects for Democratic Design," *British Journal of Politics and International Relations* 20, no. 3 (2018): 693–710, https://doi.org/10.1177/1369148118784756; Deborah Scott, "Diversifying the Deliberative Turn: Toward an Agonistic RRI," *Science, Technology, & Human Values* 48, no. 2 (2023): 295–318, https://doi.org/10.1177/01622439211067268.

55 Jack Stilgoe, Richard Owen, and Phil Macnaghten, "Developing a Framework for Responsible Innovation," *Research Policy* 42, no. 9 (2013): 1568–80, https://doi.org/10.1016/j.respol.2013.05.008.

56 Jeroen van Bouwel and Michiel van Oudheusden, "Participation beyond Consensus? Technology Assessments, Consensus Conferences and Democratic Modulation," *Social Epistemology* 31, no. 6 (2017): 497–513, https://doi.org/10.1080/02691728.2017.1352624.

57 Audley Genus and Andy Stirling, "Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation," *Research Policy* 47, no. 1 (2018): 61–69, https://doi.org/10.1016/j.respol.2017.09.012.

58 Eugen Octav Popa, Vincent Blok, and Renate Wesselink, "An Agonistic Approach to Technological Conflict," *Philosophy & Technology* 34 (December 2021): 717–37, https://doi.org/10.1007/s13347-020-00430-7.

59 Scott, "Diversifying the Deliberative Turn."

60 DiSalvo, "Design, Democracy and Agonistic Pluralism"; Kate Crawford, "Can an Algorithm Be Agonistic? Ten Scenes from Life in Calculated Publics," *Science, Technology, & Human Values* 41, no. 1 (2016): 77–92, https://doi.org/10.1177/0162243915589635; Mireille Hildebrandt, "Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning," *Theoretical Inquiries in Law* 20, no. 1 (2019): 83–121, https://doi.org/10.1515/til-2019-0004; Samantha Robertson and Niloufar Salehi, "What If I Don't Like any of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design," arXiv, July 13, 2020, https://doi.org/10.48550/arXiv.2007.06718; Peter T. Dunn, "Participatory Infrastructures: The Politics of Mobility Platforms," *Urban Planning* 5, no. 4 (2020): 335–46, https://doi.org/10.17645/up.v5i4.3483; Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz, "Hard Choices in Artificial Intelligence," *Artificial Intelligence* 300 (November 2021): 103555, https://doi.org/10.1016/j.artint.2021.103555; Deger Ozkaramanli, Armağan Karahanoğlu, and Peter-Paul Verbeek, "Reflecting on Design Methods and Democratic Technology Development: The Case of Dutch Covid-19 Digital Contact-Tracing Application," *She Ji: The Journal of Design, Economics, and Innovation* 8, no. 2 (2022): 244–69, https://doi.org/10.1016/j.sheji.2022.04.002.

61 DiSalvo, "Design, Democracy and Agonistic Pluralism."

62 Ozkaramanli et al., "Reflecting on Design Methods."

63 Crawford, "Can an Algorithm Be Agonistic."

64 Mireille Hildebrandt, "Law as Information in the Era of Data-Driven Agency," *Modern Law Review* 79, no. 1 (2016): 1–30, https://doi.org/10.1111/1468-2230.12165.

65 Robertson and Salehi, "What If I Don't Like any of the Choices?"

66 Dunn, "Participatory Infrastructures."

67 Dobbe et al., "Hard Choices in Artificial Intelligence."

68 Thomas Krendl Gilbert et al., "Reward Reports for Reinforcement Learning," in *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (New York: ACM, 2023), 84–130, https://doi.org/10.1145/3600211.3604698.

69 Paola Antonelli, "States of Design 03: Thinkering," *Domus*, July 4, 2011, https://www.domusweb.it/en/design/2011/07/04/states-of-design-03-thinkering.html.

70 Malcolm McCullough, *Digital Ground: Architecture, Pervasive Computing, and Environmental Knowing* (Cambridge, MA: MIT Press, 2005), 92–94.

71 Hugh Dubberly, "Why We Should Stop Describing Design as 'Problem Solving,'" Dubberly Design Office, October 20, 2022, http://www.dubberly.com/articles/why-we-should-stop-describing-design-as-problem-solving.html.

72 John Thackara, *In the Bubble: Designing in a Complex World* (Cambridge, MA: MIT Press, 2005), 7, 214.

73 Kristina Höök and Jonas Löwgren, "Characterizing Interaction Design by Its Ideals: A Discipline in Transition," *She Ji: The Journal of Design, Economics, and Innovation* 7, no. 1 (2021): 34, https://doi.org/10.1016/j.sheji.2020.12.001.

74 Jonas Löwgren, Bill Gaver, and John Bowers, "Annotated Portfolios and Other Forms of Intermediate-Level Knowledge," *Interactions* 20, no. 1 (2013): 30–34, https://doi.org/10.1145/2405716.2405725; Höök and Löwgren, "Strong Concepts."

75 Alfrink et al., "Contestable AI by Design"; Alfrink et al., "Contestable Camera Cars."

76 Željko Obrenović, "Design-Based Research: What We Learn When We Engage in Design of Interactive Systems," *Interactions* 18, no. 5 (2011): 56–59, https://doi.org/10.1145/2008176.2008189.

77 Katherine Isbister and Kristina Höök, "On Being Supple: In Search of Rigor Without Rigidity in Meeting New Design and Evaluation Challenges for HCI Practitioners," in *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York: ACM, 2009), 2233–42, https://doi.org/10.1145/1518701.1519042; Colin M. Gray, "'It's More of a Mindset Than a Method': UX Practitioners' Conception of Design Methods," in *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2016), 4044–55, https://doi.org/10.1145/2858036.2858410; Gray and Kou, "UX Practitioners' Engagement"; Nur Yildirim et al., "Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook," in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2023), article no. 356, https://doi.org/10.1145/3544548.3580900.

78 Gray and Kou, "UX Practitioners' Engagement."

79 Tufte, *Visual Explanations*.

80 Williams et al., "Creating Knowledge."

81 Katja Thoring, Roland Mueller, and Petra Badke-Schaub, "Workshops as a Research Method: Guidelines for Designing and Evaluating Artifacts through Workshops," in *Proceedings of the 53rd Hawaii International Conference on System Sciences 2020* (Hawaii: HICSS, 2020), 5036–45, https://hdl.handle.net/10125/64362; Daniela K. Rosner et al., "Out of Time, Out of Place: Reflections on Design Workshops as a Research Method," in *CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (New York: ACM, 2016), 1131–41, https://doi.org/10.1145/2818048.2820021.

82 Virginia Braun and Victoria Clarke, *Thematic Analysis: A Practical Guide* (Thousand Oaks, CA: SAGE, 2021).

83 Tom Fryer, "A Critical Realist Approach to Thematic Analysis: Producing Causal Explanations," *Journal of Critical Realism* 21, no. 4 (2022): 365–84, https://doi.org/10.1080/14767430.2022.2076776.

84 Bill Gaver and John Bowers, "Annotated Portfolios," *Interactions* 19, no. 4 (2012): 40–49, https://doi.org/10.1145/2212877.2212889; Löwgren et al., "Annotated Portfolios and Other Forms."

not easily predicted but arrived at iteratively over time based on human judgment.[70] In this context, designers become like stewards; their role is never finished.[71] They become facilitators of change among various stakeholders, helping them to act more intelligently in a more design-minded way in our systems.[72] As Kristina Höök and Jonas 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."[73]

Although we evaluate the Arena metaphor in the context of early-stage concept design, we do not intend its applicability 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, steering choices towards those that increase the contestability of AI systems.

Method

We aimed to develop and evaluate generative intermediate-level design knowledge.[74] This type of knowledge occupies the middle ground between specific instances and general theory, providing seeds for design solutions without prescribing their shape. We built on our prior work that introduced a framework for contestable AI.[75] Frameworks are a type of design knowledge that outlines solution characteristics for achieving goals in specific contexts.[76] We evaluated this approach with practitioners[77] to strengthen the HCI research-practitioner relationship.[78] We translated the framework and the accompanying generative metaphor of the Arena into a visual explanation.[79] Such infographics are suitable for depicting systems-oriented knowledge and are especially beneficial for practitioners who often rely on visual aids.[80] We conducted workshops with professional designers to assess the infographic, a standard method in HCI design research.[81] Our qualitative analysis of workshop outcomes used the theory of generative metaphor as a lens and employed reflexive thematic analysis,[82] further adapted using critical realist approaches[83] and annotated portfolios.[84]

Preregistration

We preregistered this study at Open Science Framework (OSF).[85] 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.

Visual Explanation Design Process

The process of constructing the visual explanation unfolded as follows. First, we drafted a creative brief. The two critical ingredients for the infographic are the Features section of the Contestable AI by Design framework[86] (Figure 1), updated with insights from the Five Loops model[87] (Figure 2), and the Agonistic Arena generative metaphor. We used the infographic loops to establish the new relations between stakeholders, an essential element of the Arena. The infographic under construction is specific to the public sector context by

85 Kars Alfrink et al., "Envisioning Contestability Loops," Open Science Framework, March 21, 2023, https://doi.org/10.17605/OSF.IO/QJZGV.
86 Alfrink et al., "Contestable AI by Design."
87 Alfrink et al., "Contestable Camera Cars."

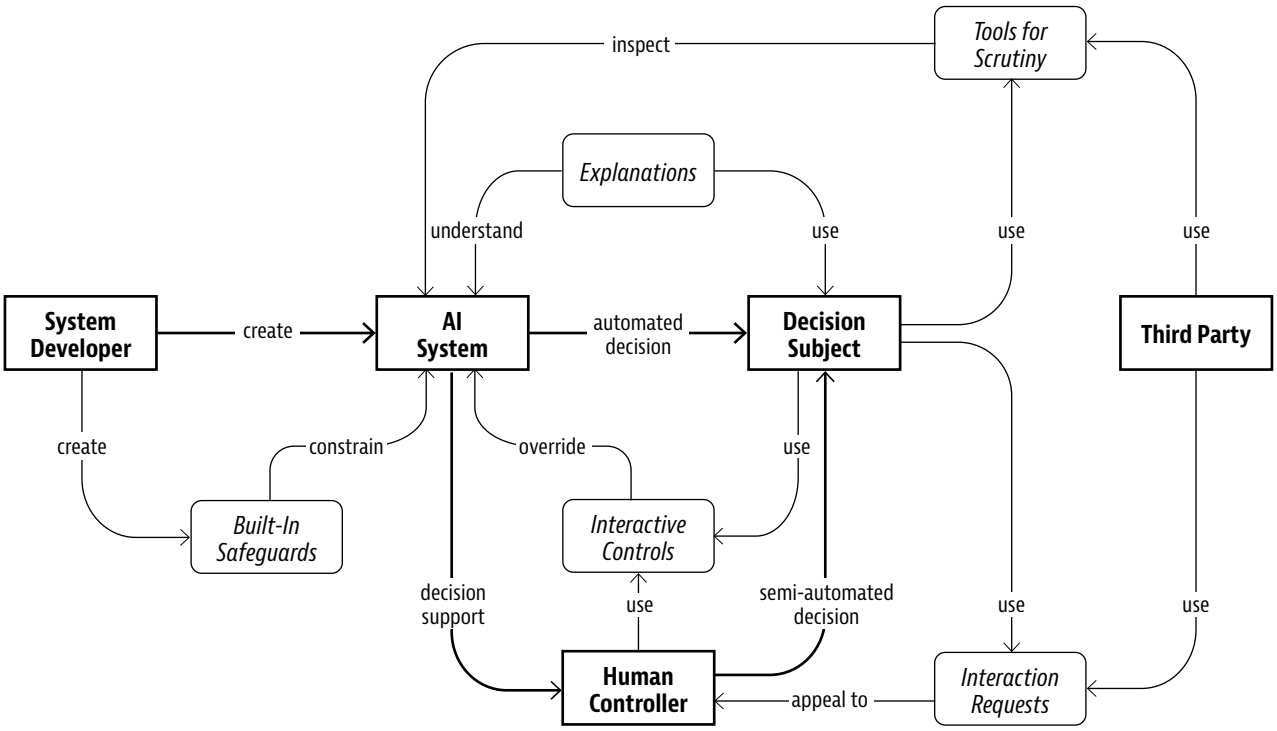


Figure 1
Features contributing to contestable AI: System developers create built-in safeguards to constrain the behavior of AI systems. Human controllers use interactive controls to correct or override AI system decisions. Decision subjects use interactive controls, explanations, intervention requests, 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. © 2022 Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn.

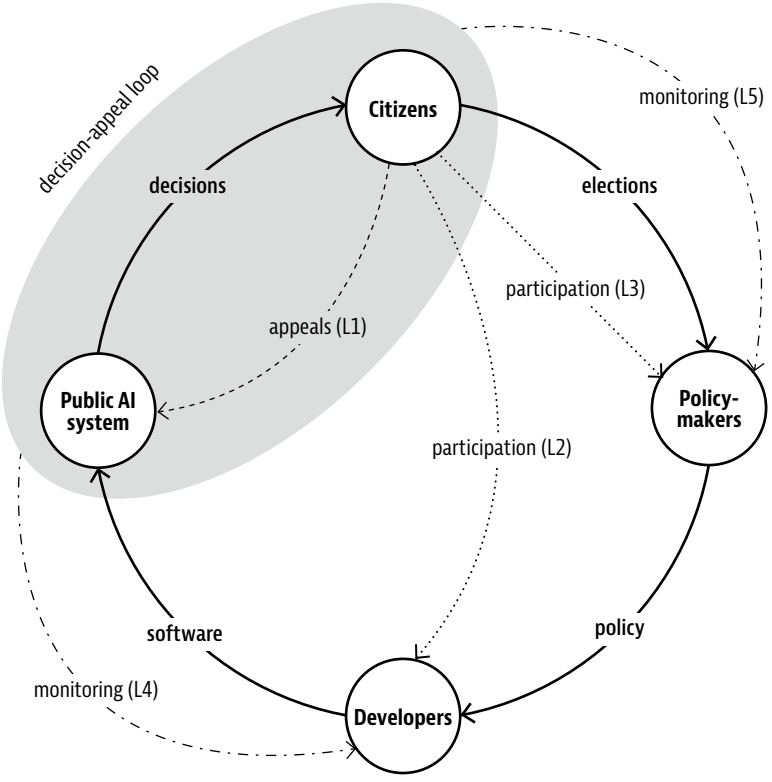


Figure 2
Diagram of "five loops model," showing the basic flow of policy through software into decisions (solid arrows), the direct way citizens can contest individual decisions (L1, dashed arrow), the direct ways in which citizens can contest systems development and policy making (L2-3, dotted arrows), 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). © 2023 Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem.

Table 1 **Definitions used in the design workshop.**

Visual explanation	"*Pictures of verbs*, the representation of mechanism and motion, of process and dynamics, of causes and effects, of explanation and narrative."[i] 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.[ii]
Design concept	Portrayals of future designs,[iii] 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."[iv]
Public AI	AI used by public sector actors to support, augment, or automate decisions.[v]
Contestable AI	Open and responsive to human intervention throughout the system lifecycle, establishing a dialectical relationship between decision subjects and system operators.[vi]

[i] Tufte, Visual Explanations.
[ii] Höök and Löwgren, "Strong Concepts."
[iii] Stolterman and Wiberg, "Concept-Driven Interaction Design Research."
[iv] Suchman, "Corporate Accountability."
[v] Nouws, Janssen, and Dobbe, "Dismantling Digital Cages."
[vi] Alfrink et al., "Contestable AI by Design."

explicitly including the representative democratic policy-making process. We aimed to develop the infographic for design practitioners, offering them more concrete guidance than the underlying theoretical framework. In the creative brief, we also delineated the key concepts mobilized in this study (Table 1).

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 underwent eight iterations between April 11 and May 22, 2023.

Throughout the process, we made some critical design decisions, 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.

Visual Explanation

The infographic depicts a generic human-AI decision-making system,[88] 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 3). The four features are *interactive controls*, *intervention requests*, *tools for scrutiny*, and *monitoring*. Interactive controls allow human

88 **A detailed anatomy of the visual explanation is provided in Appendix A.**

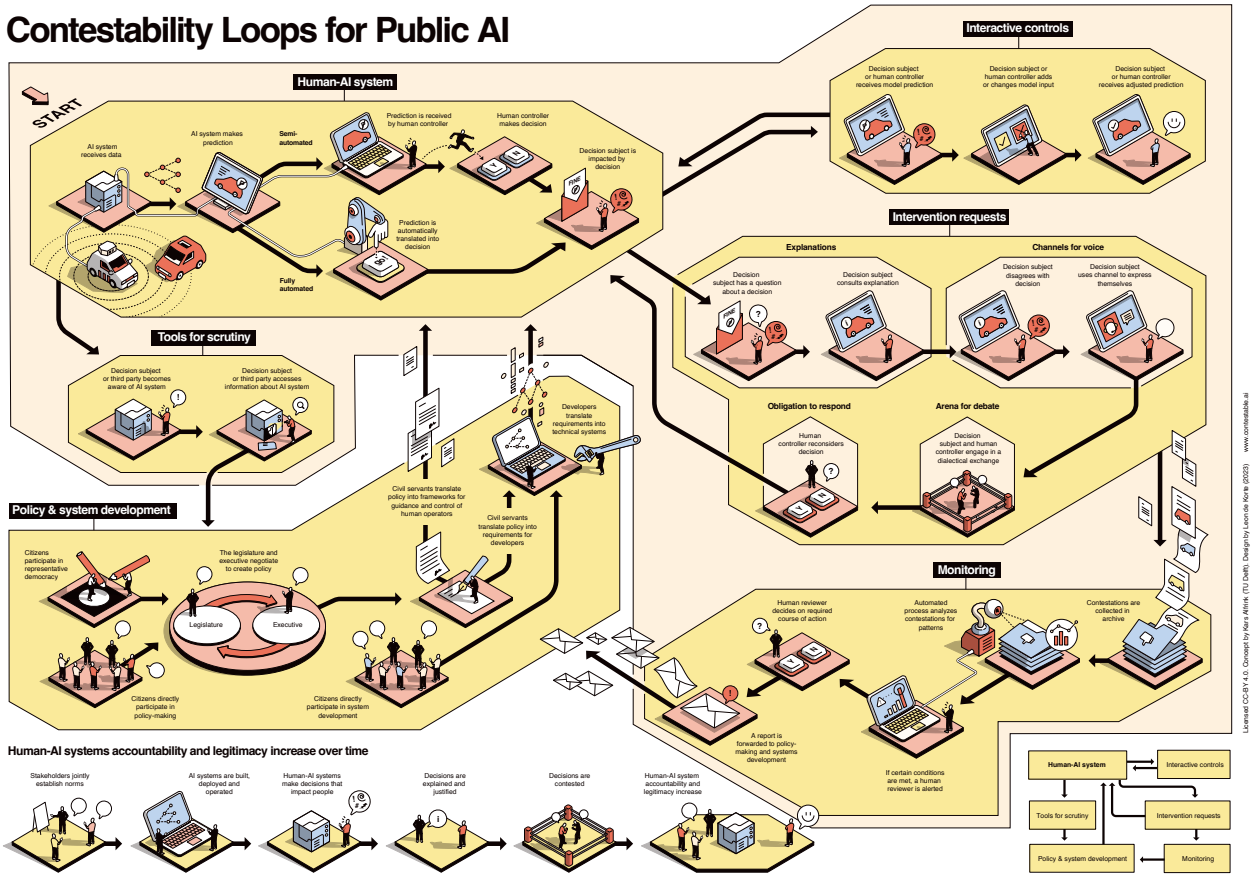


Figure 3
Contestability Loops for Public AI infographic used in workshops. © 2024 the authors.

controllers and decision subjects to intervene in the AI prediction process. Intervention requests enable data subjects to understand individual decisions, express their disagreement, debate system operators, and receive a human review of a decision. Tools for scrutiny allow a wide range of groups in society to inspect the workings of human-AI systems. Finally, Monitoring is 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, including 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 toward a more legitimate state over time under pressure from repeated contestations.

Design Workshops

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 “Contestable AI” and

"Visual Explanations." 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 (Section "Case"). 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. We provided a set of materials used to sketch; these materials were consistent across workshops.[89] 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 author was the workshop facilitator, lecturer, and guide in the final discussion. We did not actively participate in concept design exercises.

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 Virginia Braun and Victoria Clarke.[90] Workshops lasted three hours and took place on participant agencies' premises. Participants spent 33–55 minutes sketching (M = 40, SD = 11). Focus group discussions lasted 39–51 minutes (M = 44, SD = 4). The data generated consists of concept design sketches and verbal descriptions. Ten concepts were generated in total.

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

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).[91]

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 Amsterdam to screen for a system that uses risk scoring, which has become a widespread practice implicated in more than a few public administration scandals in recent history.[92] We searched for a system that addressed a relatable issue involving some stakes but was not

89 A3 marker pad, HB pencils, Sharpie markers, and Post-it notes.

90 Virginia Braun and Victoria Clarke, *Successful Qualitative Research: A Practical Guide for Beginners* (Los Angeles: SAGE, 2013), 115.

91 On a scale of one to five, one being "not at all" and five "extremely knowledgeable."

92 For example, see Adamantia Rachovitsa and Niclas Johann, "The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case," *Human Rights Law Review* 22, no. 2 (2022): ngac010, https://doi.org/10.1093/hrlr/ngac010; Marvin van Bekkum and Frederik Zuiderveen Borgesius, "Digital Welfare Fraud Detection and the Dutch SyRI Judgment," *European Journal of Social Security* 23, no. 4 (2021): 323–40, https://doi.org/10.1177/13882627211031257; Sonja Bekker, "Fundamental Rights in Digital Welfare States: The Case of SyRI in the Netherlands," in *Netherlands Yearbook of International Law 2019: Yearbooks in International Law: History, Function and Future*, ed. Otto Spijkers, Wouter G. Werner, and Ramses A. Wessel (The Hague: T.M.C. Asser Press, 2021), 289–307, https://doi.org/10.1007/978-94-6265-403-7_24.

highly polarizing. We opted for a system that the city piloted as part of the enforcement of illegal vacation rentals.

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: 1) to ensure adequate living space availability for residents and 2) to prevent visitors from adversely affecting the city's livability.

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)[93] 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.[94]

Analysis

We based our overall analysis approach on reflexive thematic analysis.[95] We adapted the approach to our purposes, drawing inspiration from critical realist approaches to thematic analysis[96]—in particular, alternating between data-led and theory-led coding and a hierarchy of codes and themes that reflected our research question (Figure 4). We took further inspiration from the annotated portfolios approach to design knowledge construction from individual design instances.[97]

The data we worked with are transcripts of verbal descriptions of concept designs supported by sketches. We coded the transcripts for verbal

93 Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, ed. Isabelle Guyon et al. (NeurIPS Foudation, 2017), 1–10, available at https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d-76c43dfd28b67767-Paper.pdf.

94 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, accessed May 23, 2023, https://amsterdam.raadsinformatie.nl/document/12800007/1.

95 Virginia Braun and Victoria Clarke, "Can I Use TA? Should I Use TA? Should I Not Use TA? Comparing Reflexive Thematic Analysis and Other Pattern-Based Qualitative Analytic Approaches," *Counselling and Psychotherapy Research* 21, no. 1 (2021): 37–47, https://doi.org/10.1002/capr.12360; Virginia Braun and Victoria Clarke, "Conceptual and Design Thinking for Thematic Analysis," *Qualitative Psychology* 9, no. 1 (2022): 3–26, https://doi.org/10.1037/qup0000196; Virginia Braun and Victoria Clarke, "Reflecting on Reflexive Thematic Analysis," *Qualitative Research in Sport, Exercise and Health* 11, no. 4 (2019): 589–97, https://doi.org/10.1080/2159676X.2019.1628806; Braun and Clarke, *Thematic Analysis*; Virginia Braun and Victoria Clarke, "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology* 3, no. 2 (2006): 77–101, https://doi.org/10.1191/1478088706qp063oa; Virginia Braun and Victoria Clarke, "Thematic Analysis," in *Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, vol. 2 of *APA Handbook of Research Methods in Psychology*, ed. Harris Cooper et al. (Washington, DC: American

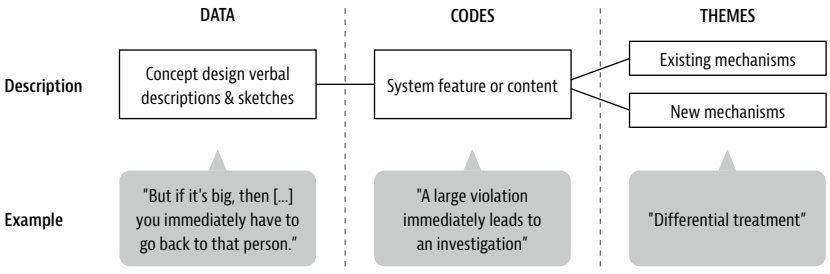


Figure 4
Conceptual model of thematic analysis of verbal concept design descriptions and accompanying sketches. © 2024 the authors.

Psychological Association, 2012), 57–71, https://doi.org/10.1037/13620-004.

96 Fryer, "Critical Realist Approach"; Gareth Wiltshire and Noora Ronkainen, "A Realist Approach to Thematic Analysis: Making Sense of Qualitative Data Through Experiential, Inferential and Dispositional Themes," *Journal of Critical Realism* 20, no. 2 (2021): 159–80, https://doi.org/10.1080/14767430.2021.1894909.

97 John Bowers, "The Logic of Annotated Portfolios: Communicating the Value of 'Research Through Design,'" in *DIS '12: Proceedings of the Designing Interactive Systems Conference* (New York: ACM, 2012), 68–77, https://doi.org/10.1145/2317956.2317968; Gaver and Bowers, "Annotated Portfolios"; Löwgren et al., "Annotated Portfolios and Other Forms."

98 For more information, see https://github.com/openai/whisper.

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

statements that refer to system features or contents. These codes were grouped into higher-level themes, each representing a single mechanism for contestability. These mechanisms were compared to the infographic to determine whether they were new or existing.

The first author performed the data analysis. The remaining authors contributed with partial coding and the review of coding results.

Data Preparation

We scanned sketches and stored them as image files to prepare the data. Then, focus group audio recordings were machine-transcribed using Whisper.[98] The first author manually edited the raw transcriptions, removed identifying 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 here.[99]

Thematic Analysis

The first author coded transcripts in Atlas.ti following the conceptual model outlined in Figure 4. 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 by the infographic.

We considered a mechanism as *existing* if it resembled an infographic feature. Mechanisms that did not resemble the infographic were deemed *new*. Throughout this process, we referred to the concept design sketches to contextualize the analysis.

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 reflected participants' intentions.

Positionality

As advocates of contestability, we aim for contestable AI to be an aspect of practice. The participants in this study are peers in the design field, including some with whom we have worked previously. These participants are employees of design agencies, some with whom we maintain professional relationships. The case is from Amsterdam, a municipality we have worked with on other studies in the past.

Results: Concept Design Mechanisms

Participants generated a total of ten concept designs. Concept descriptions are summarized in Table 2. Figure 5 shows examples of concept design sketches produced by participants. From these designs, we construct existing and new mechanisms. We summarize these results in Tables 3 and 4. 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.[100]

100 Concept designs are referred to with a C followed by a number (for example, "C2" is the concept generated in workshop two). If a workshop produced more than one concept, we gave it a suffix (for example, "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 individual participant (for example, "W1P1" is participant one in workshop one).

Table 2 **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	A system that 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.

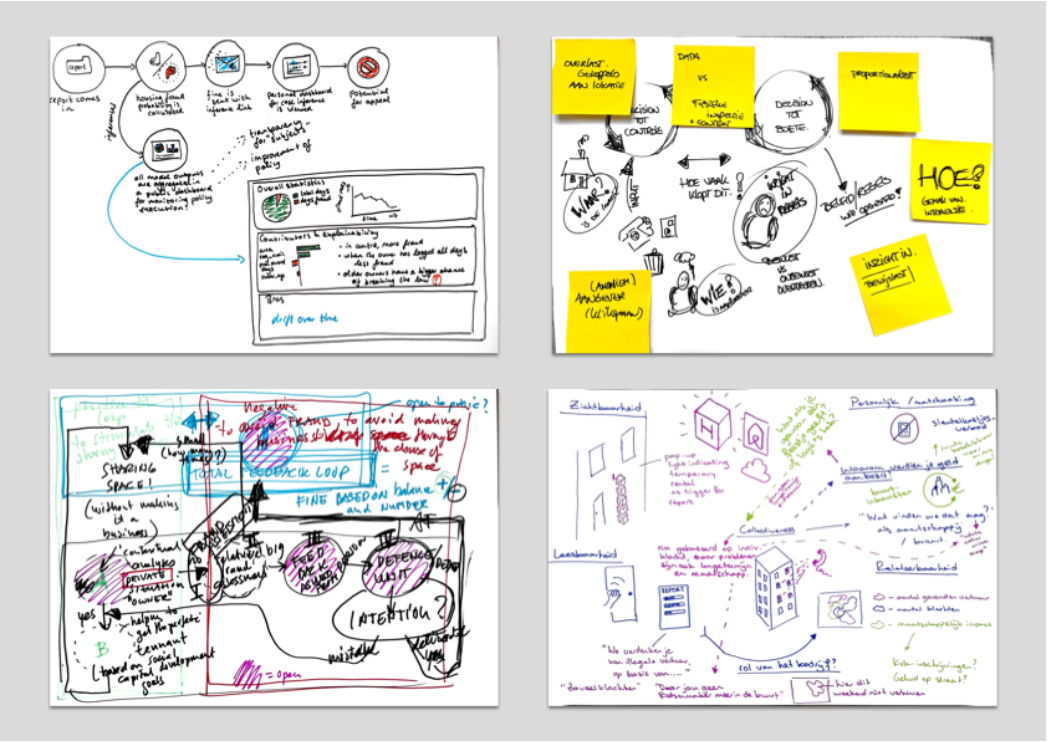


Figure 5
Examples of concept design sketches created by participants during workshops. Images courtesy of the workshop participants.

Table 3 **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.

Existing Mechanisms

The existing mechanisms that feature most prominently include 1) explanations, 2) interactive controls, 3) intervention requests, and 4) tools for scrutiny (Table 3).

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

An explanation is also included where the result was 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 (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 AI systems because they use their outputs. Global-level explanations for this purpose can be technical, but they should not be overly technical (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 was provided using SHAP.

The system should show the level of uncertainty 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). Once a report has been submitted and the AI system has produced a risk score decision, subjects should be notified immediately (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 that the decision subject can inspect immediately. If the subject indicates they disagree with the decision, an objection procedure can be initiated (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 (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 rather than 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)

New Mechanisms

The most prominent new mechanisms include 1) annual assessments, 2) differential treatment, 3) input data revisions, 4) pro-active notifications, and 5) pro-social behavior incentives (Table 4).

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 you have to pay. And you know

Table 4 **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.

> 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 through 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 aid of the AI system 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-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 matter 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 and instructions on contesting the decision (C5.1, C1.1, C3).

Several concepts include convenient touchpoints for indicating disagreement in the real world, such as 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)

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.

While vacation rentals have collective impacts, individual complaints drive the enforcement policy. Hence, 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).

> "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)

Discussion

Summary of Findings

We analyzed ten concept designs and constructed mechanisms from them 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. Their relationship to the example case human-AI system and the three generative metaphors we have constructed are shown in Figure 6.

Table 5 **Summary of metaphors, mechanisms, and concepts.**

Metaphor	Mechanism	Description	Concept designs
Agonistic Arena	*Differential treatment*	Implementing 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
	Interactive controls	Providing means for human controllers to review, adjust, and provide feedback on risk scores; providing means for citizens to respond to report contents and correct input data.	C2, C4.2, C5.1, C5.2
	Intervention requests	Providing means 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 that they are under scrutiny at every step of the process.	C1.1, C3, C4.3, C5.1
	Tools for scrutiny	Integrating AI system details into rental platforms. Providing 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*	Conducting risk scoring for all citizens every year and proactively informing them of their profile should a report be filed.	C1.1, C3
	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

Note: New mechanisms are italicized.

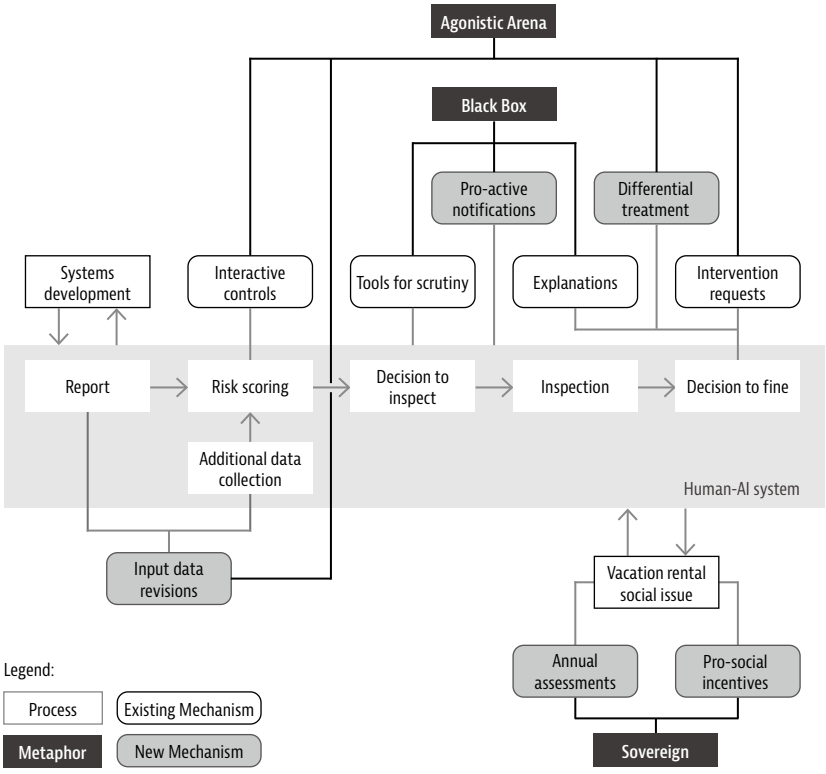


Figure 6
Diagram summarizing findings. The example case of the human-AI system process (square boxes with white fill) is related to the existing and new mechanisms proposed by the concept designs (rounded corners, white and grey fill, respectively), which in turn are related to the three generative metaphors (dark grey fill). © 2024 the authors.

In the next section, we will answer our main research question: What is the efficacy of the Agonistic Arena as a generative metaphor for the design of public AI?

Public AI as Agonistic Arena—Beyond Agreeing to Disagree

The Agonistic Arena frames public AI as a space where we celebrate all forms of struggle 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 socio-technical arrangements prioritizing mutability and reversibility.

Interactive controls and *intervention requests* are existing mechanisms that express the Agonistic Arena. *Differential treatment* and *input data revisions* are new mechanisms that do the same.

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 helpful signal for monitoring. Interactive controls enable citizens to make alternative calculations of themselves. *Intervention requests* are a necessary component of any contestable AI system, so it is no surprise that almost all concept designs include this in some form. It enables the contestation of individual decisions. Some concept designs devote more attention to the discursive element, preventing appeals from

becoming a one-way expression of discontent rather than 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* make subject to contestation the 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.

Two out of four existing 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 the AI problem in terms aligned with agonistic pluralism: a need for more discursive relationality, contestability, and contingency.

We do not consider the remaining mechanisms to 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 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.[101] 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. Further research is needed to ascertain if these metaphors extend to a broader range of design settings.

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*. *Annual assessments*, *pro-social behavior incentives*, and *pro-active notifications* are new mechanisms that do the same. We see two competing metaphors in the design space covered by the concept designs: the Black Box — AI as an opaque system that requires opening up — and the Sovereign — AI as an all-knowing 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.[102] This opacity can stem from secrecy, illiteracy, or scale and complexity.[103] The Black Box metaphor is central to the field of explainable AI (XAI),[104] which develops technical solutions to the fundamental opacity of ML models.

101 Philippe Lorino, "Abduction," in *Pragmatism and Organization Studies* (Oxford: Oxford University Press, 2018), 189–222, https://doi.org/10.1093/oso/9780198753216.003.0007.

102 Jonathan A. Obar, "Sunlight alone Is Not a Disinfectant: Consent and the Futility of Opening Big Data Black Boxes (Without Assistance)," *Big Data & Society* 7, no. 1 (2020): 1–5, https://doi.org/10.1177/2053951720935615.

103 Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3, no. 1 (2016): 1–12, https://doi.org/10.1177/2053951715622512.

104 Alejandro Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* 58 (June 2020): 82–115, https://doi.org/10.1016/j.inffus.2019.12.012; Amina Adadi and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* 6 (2018): 52138–60, https://doi.org/gfvb5g.

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 need to be revised for contestability because they lack the normative dimension; they do not justify[105] why a decision is desirable. Without justification, a decision subject cannot mount an "articulate act of defense."[106] 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 how it works 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,[107] enabling operators to exercise discretion in applying decision rules.[108] Additionally, it proposes a sociotechnical approach, focusing on collective understanding and dialogue[109] rather than individual interpretation, to overcome limitations in fully explaining machine outputs.[110] This approach critiques the Black Box for neglecting power dynamics and unrealistically assuming liberal ideals of free and equal individuals.

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

The Sovereign: Save Us from Ourselves

The *annual assessments* and *pro-social behavior incentives* are new mechanisms that express a less prominent but intriguing metaphor—the Sovereign. This metaphor 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.

105 Henin and Le Métayer, "Beyond Explainability."
106 Sarra, "Put Dialectics into the Machine."
107 Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1, no. 5 (2019): 206–15, https://doi.org/10.1038/s42256-019-0048-x.
108 Ali Alkhatib and Michael Bernstein, "Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions," in *CHI '19: Conference on Human Factors in Computing Systems* (New York: ACM, 2019), paper no. 530, https://doi.org/10.1145/3290605.3300760; Mark Bovens and Stavros Zouridis, "From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control," *Public Administration Review* 62, no. 2 (2002): 174–84, https://doi.org/10.1111/0033-3352.00168; Marie Leth Meilvang and Anne Marie Dahler, "Decision Support and Algorithmic Support: The Construction of Algorithms and Professional Discretion in Social Work," *European Journal of Social Work* 27, no. 1 (2022): 30–42, https://doi.org/10.1080/13691457.2022.2063806; Anette C. M. Petersen, Lars Rune Christensen, and Thomas T. Hildebrandt, "The Role of Discretion in the Age of Automation," *Computer Supported Cooperative Work* 29, no. 3 (2020): 303–33, https://doi.org/10.1007/s10606-020-09371-3.
109 Mike Ananny and Kate Crawford, "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media and Society* 20, no. 3 (2018): 973–89, https://doi.org/10.1177/1461444816676645.
110 Joel Walmsley, "Artificial Intelligence and the Value of Transparency," *AI & SOCIETY* 36, no. 2 (2021): 585–95, https://doi.org/10.1007/s00146-020-01066-z.

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 extreme.[111] *Pro-social behavior incentives* use 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 to an autocratic data-driven apparatus.

The issue of authoritarian AI is a longstanding concern in critical AI studies. Typically, people perceive algorithms 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.[112] Concerns over the imposition of data-driven cybernetic choice architectures are also enduring in critical smart cities research.[113]

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

Relationships between Arena, Black Box, and Sovereign

Here, we briefly examine the relationships between the three metaphors, drawing on Alex Hochuli, George Hoare, and Philip Cunliffe's framework of politics, post-politics, and antipolitics.[114]

The Black Box metaphor represents a postpolitical stance, implying that resolving public AI issues merely requires providing more information, overlooking 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 the inadequacies of this order and, ironically, suggests eliminating politics with the assistance of AI. The Arena metaphor acknowledges similar frustrations but advocates for further democratization of AI and emphasizes political contestation.[115]

111 Matthew M. Young, Justin B. Bullock, and Jesse D. Lecy, "Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration," *Perspectives on Public Management and Governance* 2, no. 4 (2019): 301–13, https://doi.org/10.1093/ppmgov/gvz014; Patrick Dunleavy et al., "New Public Management Is Dead — Long Live Digital-Era Governance," *Journal of Public Administration Research and Theory* 16, no. 3 (2006): 467–94, https://doi.org/10.1093/jopart/mui057.

112 Crawford, "Can an Algorithm Be Agonistic," 79, 86–87.

113 Dan McQuillan, "People's Councils for Ethical Machine Learning," *Social Media + Society* 4, no. 2 (2018): online, https://doi.org/10.1177/2056305118768303; Taylor Shelton, Matthew Zook, and Alan Wiig, "The 'Actually Existing Smart City,'" *Cambridge Journal of Regions, Economy and Society* 8, no. 1 (2015): 13–25, https://doi.org/10.1093/cjres/rsu026; Rob Kitchin, "The Real-Time City? Big Data and Smart Urbanism," *GeoJournal* 79, no. 1 (2014): 1–14, https://doi.org/10.1007/s10708-013-9516-8; Jathan Sadowski and Frank Pasquale, "The Spectrum of Control: A Social Theory of the Smart City," *First Monday* 20, no. 7 (2015): 1–22, https://doi.org/10.5210/fm.v20i7.5903; Marcus Foth, "Participatory Urban Informatics: Towards Citizen-Ability," *Smart and Sustainable Built Environment* 7, no. 1 (2018): 4–19, https://doi.org/10.1108/SASBE-10-2017-0051.

114 Alex Hochuli, George Hoare, and Philip Cunliffe, *The End of the End of History: Politics in the Twenty-First Century* (Ridgefield, CT: Zero Books, 2021), 13.

115 John W. Murphy and Randon R. Taylor, "To Democratize or Not to Democratize AI? That Is the Question," *AI and Ethics*, June 15, 2023, https://doi.org/10.1007/s43681-023-00313-5.

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. 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 metaphors. Moreover, crafting new metaphors will be necessary if we seek to change the way designers frame problems related to AI. Such metaphors can be assembled from theory, as is the case in concept-driven interaction design research,[116] 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. While clarifying the central metaphor can help designers better understand tools and techniques, tool creators can never reliably transfer intent entirely. For example, some of our participants adapted tools for contestability in a manner more in line with the Black Box metaphor, which is popular in HCI design. This discrepancy may stem from our participant designers' preference for 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.[117] Even designers focusing on human-centered AI in public settings can hold conflicting views. Without a thorough grasp of the political philosophies influencing our designs and articulating a coherent stance, we run the risk of developing inconsistent proposals.

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 rich 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 the contestability of public AI systems. 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 reflected 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,[118] and since using particular metaphors to frame design challenges leads to particular diagnoses and accompanying prescriptions,[119] design research and practice are well-served by

116 Erik Stolterman and Mikael Wiberg, "Concept-Driven Interaction Design Research," *Human-Computer Interaction* 25, no. 2 (2010): 95–118, https://doi.org/10.1080/07370020903586696.

117 John T. Jost, Christopher M. Federico, and Jaime L. Napier, "Political Ideology: Its Structure, Functions, and Elective Affinities," *Annual Review of Psychology* 60, no. 1 (2009): 307–37, https://doi.org/10.1146/annurev.psych.60.110707.163600.

118 Lakoff, *Women, Fire, and Dangerous Things*.

119 Schön, "Generative Metaphor."

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.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

While preparing this work, the authors used ChatGPT and GrammarlyGO to edit self-written text for clarity and brevity. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Declaration of Interest

There are no conflicts of interest involved in this article.

Acknowledgments

We thank Responsible Sensing Lab for funding the infographic production and general support; Thijs Turel for all his support; Rob Collins 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. This research was supported by a grant from the Dutch National Research Council NWO (grant no. CISC.CC.018).

References

Adadi, Amina, and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6 (2018): 52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Alfrink, Kars, Ianus Keller, Neelke Doorn, and Gerd Kortuem. "Tensions in Transparent Urban AI: Designing a Smart Electric Vehicle Charge Point." *AI & Society* 3 (June 2022): 1049–65. https://doi.org/10.1007/s00146-022-01436-9.

Alfrink, Kars, Ianus Keller, Gerd Kortuem, and Neelke Doorn. "Contestable AI by Design: Towards a Framework." *Minds and Machines* 33 (August 2022): 613–39. https://doi.org/10.1007/s11023-022-09611-z.

Alfrink, Kars, Ianus Keller, Neelke Doorn, and Gerd Kortuem. "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*, article no. 8. New York: ACM, 2023. https://doi.org/10.1145/3544548.3580984.

Alfrink, Kars, Ianus Keller, Gerd Kortuem, and Neelke Doorn. "Envisioning Contestability Loops." Open Science Framework, March 21, 2023. https://doi.org/10.17605/OSF.IO/QJZGV.

Alkhatib, Ali, and Michael Bernstein. "Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions." In *CHI '19: Conference on Human Factors in Computing Systems*, paper no. 530. New York: ACM, 2019. https://doi.org/10.1145/3290605.3300760.

Almada, Marco. "Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems." In *ICAIL '19: Proceedings of the 17th International Conference on Artificial Intelligence and Law*, 2–11. New York: ACM, 2019. https://doi.org/10.1145/3322640.3326699.

Ananny, Mike, and Kate Crawford. "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media and Society* 20, no. 3 (2018): 973–89. https://doi.org/10.1177/1461444816676645.

Antonelli, Paola. "States of Design 03: Thinkering." *Domus*, July 4, 2011. https://www.domusweb.it/en/design/2011/07/04/states-of-design-03-thinkering.html.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58 (June 2020): 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Beck, Jordan, and Hamid R. Ekbia. "The Theory-Practice Gap as Generative Metaphor." In *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, paper no. 620. New York: ACM, 2018. https://doi.org/10.1145/3173574.3174194.

Bekker, Sonja. "Fundamental Rights in Digital Welfare States: The Case of SyRI in the Netherlands." In *Netherlands Yearbook of International Law 2019: Yearbooks in International Law: History, Function and Future*, edited by Otto Spijkers, Wouter G. Werner, and Ramses A. Wessel, 289–307. The Hague: T.M.C. Asser Press, 2021. https://doi.org/10.1007/978-94-6265-403-7_24.

Benjamin, Jesse Josua, Heidi Biggs, Arne Berger, Julija Rukanskaitė, Michael B. Heidt, Nick Merrill, James Pierce, and Joseph Lindley. "The Entoptic Field Camera as Metaphor-Driven Research-Through-Design with AI Technologies." In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, article no. 178. New York: ACM, 2023. https://doi.org/10.1145/3544548.3581175.

Bovens, Mark, and Stavros Zouridis. "From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control." *Public Administration Review* 62, no. 2 (2002): 174–84. https://doi.org/10.1111/0033-3352.00168.

Bowers, John. "The Logic of Annotated Portfolios: Communicating the Value of 'Research Through Design.'" In *DIS '12: Proceedings of the Designing Interactive Systems Conference*, 68–77. New York: ACM, 2012. https://doi.org/10.1145/2317956.2317968.

Braun, Virginia, and Victoria Clarke. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3, no. 2 (2006): 77–101. https://doi.org/10.1191/1478088706qp063oa.

Braun, Virginia, and Victoria Clarke. "Thematic Analysis." In *Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, vol. 2 of *APA Handbook of Research Methods in Psychology*, edited by Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher, 57–71. Washington, DC: American Psychological Association, 2012. https://doi.org/10.1037/13620-004.

Braun, Virginia, and Victoria Clarke. *Successful Qualitative Research: A Practical Guide for Beginners*. Los Angeles: SAGE, 2013.

Braun, Virginia, and Victoria Clarke. "Reflecting on Reflexive Thematic Analysis." *Qualitative Research in Sport, Exercise and Health* 11, no. 4 (2019): 589–97. https://doi.org/10.1080/2159676X.2019.1628806.

Braun, Virginia, and Victoria Clarke. *Thematic Analysis: A Practical Guide*. Thousand Oaks, CA: SAGE, 2021.

Braun, Virginia, and Victoria Clarke. “Can I Use TA? Should I Use TA? Should I Not Use TA? Comparing Reflexive Thematic Analysis and Other Pattern-Based Qualitative Analytic Approaches.” *Counselling and Psychotherapy Research* 21, no. 1 (2021): 37–47. https://doi.org/10.1002/capr.12360.

Braun, Virginia, and Victoria Clarke. “Conceptual and Design Thinking for Thematic Analysis.” *Qualitative Psychology* 9, no. 1 (2022): 3–26. https://doi.org/10.1037/qup0000196.

Brown, Anna, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. “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*, paper no. 41. New York: ACM, 2019. https://doi.org/10.1145/3290605.3300271.

Burrell, Jenna. “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms.” *Big Data & Society* 3, no. 1 (2016): 1–12. https://doi.org/10.1177/2053951715622512.

Capel, Tara, and Margot Brereton. “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*, paper no. 359. New York: ACM, 2023. https://doi.org/10.1145/3544548.3580959.

Crawford, Kate. “Can an Algorithm Be Agonistic? Ten Scenes from Life in Calculated Publics.” *Science, Technology, & Human Values* 41, no. 1 (2016): 77–92. https://doi.org/10.1177/0162243915589635.

Cugurullo, Federico. “Urban Artificial Intelligence: From Automation to Autonomy in the Smart City.” *Frontiers in Sustainable Cities* 2 (July 2020): 1–14. https://doi.org/10.3389/frsc.2020.00038.

Desai, Smit, and Michael Twidale. “Metaphors in Voice User Interfaces: A Slippery Fish.” *ACM Transactions on Computer-Human Interaction* 30, no. 6 (2023): article no. 89. https://doi.org/10.1145/3609326.

DiSalvo, Carl. “Design, Democracy and Agonistic Pluralism.” In *Design and Complexity — DRS International Conference 2010*, edited by David Durling, R. Bousbaci, L. Chen, P. Gauthier, T. Poldma, S. Roworth-Stokes, and E. Stolterman. Montreal, Canada: DRS, 2010. https://dl.designresearchsociety.org/drs-conference-papers/drs2010/researchpapers/31.

Dobbe, Roel, Thomas Krendl Gilbert, and Yonatan Mintz. “Hard Choices in Artificial Intelligence.” *Artificial Intelligence* 300 (November 2021): 103555. https://doi.org/10.1016/j.artint.2021.103555;

Dove, Graham, and Anne-Laure Fayard. “Monsters, Metaphors, and Machine Learning.” In *CHI ’20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–17. New York: ACM, 2020. https://doi.org/10.1145/3313831.3376275.

Drobotowicz, Karolina, Marjo Kauppinen, and Sari Kujala. “Trustworthy AI Services in the Public Sector: What Are Citizens Saying about It?” In *Requirements Engineering: Foundation for Software Quality*, edited by Fabiano Dalpiaz and Paola Spoletini, 99–115. Cham: Springer, 2021. https://doi.org/10.1007/978-3-030-73128-1_7.

Dubberly, Hugh. “Why We Should Stop Describing Design as ‘Problem Solving.’” Dubberly Design Office, October 20, 2022. http://www.dubberly.com/articles/why-we-should-stop-describing-design-as-problem-solving.html.

Dunleavy, Patrick, Helen Margetts, Simon Bastow, and Jane Tinkler. “New Public Management Is Dead — Long Live Digital-Era Governance.” *Journal of Public Administration Research and Theory* 16, no. 3 (2006): 467–94. https://doi.org/10.1093/jopart/mui057.

Dunn, Peter T. “Participatory Infrastructures: The Politics of Mobility Platforms.” *Urban Planning* 5, no. 4 (2020): 335–46. https://doi.org/10.17645/up.v5i4.3483.

Fabrocini, Filippo, and Kostas Terzidis. "Re-framing AI: An AI Product Designer Perspective." *Techné: Research in Philosophy and Technology* 25, no. 3 (2021): 407–33. https://doi.org/10.5840/techne2021127151.

Fatima, Samar, Kevin C. Desouza, Christoph Buck, and Erwin Fielt. "Public AI Canvas for AI-enabled Public Value: A Design Science Approach." *Government Information Quarterly* 39, no. 4 (2022): 101722. https://doi.org/10.1016/j.giq.2022.101722.

Filonik, Jakub. "We Are the Champions: The Role of Agonistic Metaphor in the Political Discourse of Classical Greece." In *The* Agōn *in Classical Literature: Studies in Honour of Professor Chris Carey*, edited by Michael Edwards, Anastasios Eustathiu, Iōanna Karamanu, Elenē Bolonakē, and Christopher Carey, 155–62. London: University of London Press, 2022. https://doi.org/10.14296/wkue3508.

Fine Licht, Karl, and Jenny de Fine Licht. "Artificial Intelligence, Transparency, and Public Decision-Making: Why Explanations Are Key When Trying to Produce Perceived Legitimacy." *AI & SOCIETY* 35 (December 2020): 917–26. https://doi.org/10.1007/s00146-020-00960-w.

Flügge, Asbjørn Ammitzbøll. "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*, 253–55. New York: ACM, 2021. https://doi.org/10.1145/3462204.3481787.

Foth, Marcus. "Participatory Urban Informatics: Towards Citizen-Ability." *Smart and Sustainable Built Environment* 7, no. 1 (2018): 4–19. https://doi.org/10.1108/SASBE-10-2017-0051.

Frauenberger, Christopher. "Critical Realist HCI." In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 341–51. New York: ACM, 2016. https://doi.org/10.1145/2851581.2892569.

Fryer, Tom. "A Critical Realist Approach to Thematic Analysis: Producing Causal Explanations." *Journal of Critical Realism* 21, no. 4 (2022): 365–84. https://doi.org/10.1080/14767430.2022.2076776.

Gaver, Bill, and John Bowers. "Annotated Portfolios." *Interactions* 19, no. 4 (2012): 40–49. https://doi.org/10.1145/2212877.2212889.

Genus, Audley, and Andy Stirling. "Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation." *Research Policy* 47, no. 1 (2018): 61–69. https://doi.org/10.1016/j.respol.2017.09.012.

Gilbert, Thomas Krendl, Nathan Lambert, Sarah Dean, Tom Zick, Aaron Snoswell, and Soham Mehta. "Reward Reports for Reinforcement Learning." In *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 84–130. New York: ACM, 2023. https://doi.org/10.1145/3600211.3604698.

Gorski, Philip S. "'What Is Critical Realism? And Why Should You Care?'" *Contemporary Sociology: A Journal of Reviews* 42, no. 5 (2013): 658–70. https://doi.org/10.1177/0094306113499533.

Gray, Colin M. "'It's More of a Mindset Than a Method': UX Practitioners' Conception of Design Methods." In *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4044–55. New York: ACM, 2016. https://doi.org/10.1145/2858036.2858410.

Gray, Colin M., and Yubo Kou. "UX Practitioners' Engagement with Intermediate-Level Knowledge." In *DIS '17 Companion: Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems*, 13–17. New York: ACM, 2017. https://doi.org/10.1145/3064857.3079110.

Green, Stuart D., Chung-Chin Kao, and Graeme D. Larsen. "Contextualist Research: Iterating between Methods While Following an Empirically Grounded Approach." *Journal of Construction Engineering and Management* 136, no. 1 (2010): 117–26. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000027.

Hamilton, Anne. "Metaphor in Theory and Practice: The Influence of Metaphors on Expectations." *ACM Journal of Computer Documentation* 24, no. 4 (2000): 237–53. https://doi.org/10.1145/353927.353935.

Harvey, David. "The Right to the City." *International Journal of Urban and Regional Research* 27, no. 4 (2003): 939–41. https://doi.org/10.1111/j.0309-1317.2003.00492.x.

Hekkert, Paul, and Nazlı Cila. "Handle with Care! Why and How Designers Make Use of Product Metaphors." *Design Studies* 40 (September 2015): 196–217. https://doi.org/10.1016/j.destud.2015.06.007.

Henin, Clément, and Daniel Le Métayer. "Beyond Explainability: Justifiability and Contestability of Algorithmic Decision Systems." *AI & SOCIETY* 37 (December 2022): 1397–1410. https://doi.org/10.1007/s00146-021-01251-8.

Hildebrandt, Mireille. "Law as Information in the Era of Data-Driven Agency." *Modern Law Review* 79, no. 1 (2016): 1–30. https://doi.org/10.1111/1468-2230.12165.

Hildebrandt, Mireille. "Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning." *Theoretical Inquiries in Law* 20, no. 1 (2019): 83–121. https://doi.org/10.1515/til-2019-0004;

Hirsch, Tad, Kritzia Merced, Shrikanth Shri Narayanan, Zac E. Imel, and David C. Atkins. "Designing Contestability: Interaction Design, Machine Learning, and Mental Health." In *DIS '17: Proceedings of the 2017 Conference on Designing Interactive Systems*, 95–99. New York: ACM, 2017. https://doi.org/10.1145/3064663.3064703.

Hochuli, Alex, George Hoare, and Philip Cunliffe. *The End of the End of History: Politics in the Twenty-First Century*. Ridgefield, CT: Zero Books, 2021.

Höök, Kristina, and Jonas Löwgren. "Strong Concepts: Intermediate-Level Knowledge in Interaction Design Research." *ACM Transactions on Computer-Human Interaction* 19, no. 3 (2012): 1–18. https://doi.org/10.1145/2362364.2362371.

Höök, Kristina, and Jonas Löwgren. "Characterizing Interaction Design by Its Ideals: A Discipline in Transition." *She Ji: The Journal of Design, Economics, and Innovation* 7, no. 1 (2021): 24–40. https://doi.org/10.1016/j.sheji.2020.12.001.

Isbister, Katherine, and Kristina Höök. "On Being Supple: In Search of Rigor Without Rigidity in Meeting New Design and Evaluation Challenges for HCI Practitioners." In *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2233–42. New York: ACM, 2009. https://doi.org/10.1145/1518701.1519042.

Jost, John T., Christopher M. Federico, and Jaime L. Napier. "Political Ideology: Its Structure, Functions, and Elective Affinities." *Annual Review of Psychology* 60, no. 1 (2009): 307–37. https://doi.org/10.1146/annurev.psych.60.110707.163600.

Kitchin, Rob. "The Real-Time City? Big Data and Smart Urbanism." *GeoJournal* 79, no. 1 (2014): 1–14. https://doi.org/10.1007/s10708-013-9516-8.

Koskinen, Ilpo, Thomas Binder, and Johan Redström. "Lab, Field, Gallery, and Beyond." *Artifact* 2, no. 1 (2008): 46–57. https://doi.org/10.1080/17493460802303333.

Koskinen, Ilpo, John Zimmerman, Thomas Binder, Johan Redström, and Stephan Wensveen. *Design Research through Practice: From the Lab, Field, and Showroom*. Boston: Morgan Kaufmann, 2012. https://doi.org/10.1016/B978-0-12-385502-2.00013-4.

Lakoff, George. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press, 1987.

Lefebvre, Henri. "Le Droit à La Ville." *L'Homme Et La Société* 6, no. 1 (1967): 29–35. Available at https://www.persee.fr/doc/homso_0018-4306_1967_num_6_1_1063.

Lindh, Maria. "As a Utility — Metaphors of Information Technologies." *Human IT: Journal for Information Technology Studies as a Human Science* 13, no. 2 (2016): 47–80. https://humanit.hb.se/article/view/418.

Logler, Nick, Daisy Yoo, and Batya Friedman. "Metaphor Cards: A How-to-Guide for Making and Using a Generative Metaphorical Design Toolkit." In *DIS '18: Proceedings of the 2018 Designing Interactive Systems Conference*, 1373–86. New York: ACM, 2018. https://doi.org/10.1145/3196709.3196811.

Lorino, Philippe. "Abduction." In *Pragmatism and Organization Studies*, 189–222. Oxford: Oxford University Press, 2018. https://doi.org/10.1093/oso/9780198753216.003.0007.

Löwgren, Jonas, Bill Gaver, and John Bowers. "Annotated Portfolios and Other Forms of Intermediate-Level Knowledge." *Interactions* 20, no. 1 (2013): 30–34. https://doi.org/10.1145/2405716.2405725.

Lowndes, Vivien, and Marie Paxton. "Can Agonism Be Institutionalised? Can Institutions Be Agonised? Prospects for Democratic Design." *British Journal of Politics and International Relations* 20, no. 3 (2018): 693–710. https://doi.org/10.1177/1369148118784756.

Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*, vol. 30, edited by Isabelle Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 1–10. NeurIPS Foudation, 2017. Available at https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Luusua, Aale, and Johanna Ylipulli. "Artificial Intelligence and Risk in Design." In *DIS '20: Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 1235–44. New York: ACM, 2020. https://doi.org/10.1145/3357236.3395491.

Luusua, Aale, and Johanna Ylipulli. "Urban AI: Formulating an Agenda for the Interdisciplinary Research of Artificial Intelligence in Cities." In *DIS' 20 Companion: Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, 373–76. New York: ACM, 2020. https://doi.org/10.1145/3393914.3395905.

Luusua, Aale, Johanna Ylipulli, Marcus Foth, and Alessandro Aurigi. "Urban AI: Understanding the Emerging Role of Artificial Intelligence in Smart Cities." *AI & SOCIETY* 38 (June 2023): 1039–44. https://doi.org/10.1007/s00146-022-01537-5.

Lyons, Henrietta, Eduardo Velloso, and Tim Miller. "Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW1 (2021): article no. 106. https://doi.org/10.1145/3449180.

Lyons, Henrietta, Tim Miller, and Eduardo Velloso. "Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review." In *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 764–74. New York: ACM, 2023. https://doi.org/10.1145/3593013.3594041.

Madill, Anna, Abbie Jordan, and Caroline Shirley. "Objectivity and Reliability in Qualitative Analysis: Realist, Contextualist and Radical Constructionist Epistemologies." *British Journal of Psychology* 91, no. 1 (2000): 1–20. https://doi.org/10.1348/000712600161646.

Marda,Vidushi, and Shivangi Narayan. "Data in New Delhi's Predictive Policing System." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 317–24. New York: ACM, 2020. https://doi.org/10.1145/3351095.3372865.

McCullough, Malcolm. *Digital Ground: Architecture, Pervasive Computing, and Environmental Knowing*. Cambridge, MA: MIT Press, 2005.

McQuillan, Dan. "People's Councils for Ethical Machine Learning." *Social Media + Society* 4, no. 2 (2018): online. https://doi.org/10.1177/2056305118768303.

Meilvang, Marie Leth, and Anne Marie Dahler. "Decision Support and Algorithmic Support: The Construction of Algorithms and Professional Discretion in Social Work." *European Journal of Social Work* 27, no. 1 (2022): 30–42. https://doi.org/10.1080/13691457.2022.2063806.

Mhlambi, Sábëlo, and Simona Tiribelli. "Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms." *Topoi* 42 (February 2023): 867–80. https://doi.org/10.1007/s11245-022-09874-2.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices." *Science and Engineering Ethics* 26 (December 2019): 2141–68. https://doi.org/10.1007/s11948-019-00165-5.

Mouffe, Chantal. *The Return of the Political*. London: Verso, 1993.

Mouffe, Chantal. "Deliberative Democracy or Agonistic Pluralism?" *Social Research* 66, no. 3 (1999): 745–758. https://www.jstor.org/stable/40971349.

Mouffe, Chantal. *The Democratic Paradox*. London: Verso, 2000.

Mouffe, Chantal. "Pluralism, Dissensus and Democratic Citizenship." In *Education and the Good Society*, edited by Fred Inglis, 42–53. New York: Springer, 2004.

Mouffe, Chantal. *On the Political*. London: Routledge, 2005.

Mouffe, Chantal. "Some Reflections on an Agonistic Approach to the Public." In *Making Things Public: Atmospheres of Democracy*, edited by Bruno Latour and Peter Weibel, 804–7. Cambridge, MA: MIT Press, 2005.

Mouffe, Chantal. *Agonistics: Thinking the World Politically*. London: Verso, 2013.

Murphy, John W., and Randon R. Taylor. "To Democratize or Not to Democratize AI? That Is the Question." *AI and Ethics*, June 15, 2023. https://doi.org/10.1007/s43681-023-00313-5.

Murray-Rust, Dave, Iohanna Nicenboim, and Dan Lockton. "Metaphors for Designers Working with AI." In *DRS Biennial Conference Series*, edited by Dan Lockton, S. Lenzi, P. Hekkert, A. Oak, J. Sádaba, P. Lloyd, 1–20. Bilbao, Spain: DRS, 2022. https://doi.org/10.21606/drs.2022.667.

Nicenboim, Iohanna, Shruthi Venkat, Neva Linn Rustad, Diana Vardanyan, Elisa Giaccardi, and Johan Redström. "Conversation Starters: How Can We Misunderstand AI Better?" In *CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, article no. 431. New York: ACM, 2023. https://doi.org/10.1145/3544549.3583914.

Nouws, Sem, Marijn Janssen, and Roel Dobbe. "Dismantling Digital Cages: Examining Design Practices for Public Algorithmic Systems." In *Electronic Government: EGOV 2022, Lecture Notes in Computer Science*, vol. 13391, edited by Marijn Janssen, Csaba Csáki, Ida Lindgren, Euripidis Loukis, Ulf Melin, Gabriela Viale Pereira, Manuel Pedro Rodríguez Bolívar, and Efthimios Tambouris, 307–22. Cham: Springer-Verlag, 2022. https://doi.org/10.1007/978-3-031-15086-9_20.

Obar, Jonathan A. "Sunlight alone Is Not a Disinfectant: Consent and the Futility of Opening Big Data Black Boxes (Without Assistance)." *Big Data & Society* 7, no. 1 (2020): 1–5. https://doi.org/10.1177/2053951720935615.

Obrenović, Željko. "Design-Based Research: What We Learn When We Engage in Design of Interactive Systems." *Interactions* 18, no. 5 (2011): 56–59. https://doi.org/10.1145/2008176.2008189.

Ozkaramanli, Deger, Armağan Karahanoğlu, and Peter-Paul Verbeek. "Reflecting on Design Methods and Democratic Technology Development: The Case of Dutch Covid-19 Digital Contact-Tracing Application." *She Ji: The Journal of Design, Economics, and Innovation* 8, no. 2 (2022): 244–69. https://doi.org/10.1016/j.sheji.2022.04.002.

Petersen, Anette C. M., Lars Rune Christensen, and Thomas T. Hildebrandt. "The Role of Discretion in the Age of Automation." *Computer Supported Cooperative Work* 29, no. 3 (2020): 303–33. https://doi.org/10.1007/s10606-020-09371-3.

Pine, Kathleen H., and Max Liboiron. "The Politics of Measurement and Action." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3147–56. New York: ACM, 2015. https://doi.org/10.1145/2702123.2702298.

Popa, Eugen Octav, Vincent Blok, and Renate Wesselink. "An Agonistic Approach to Technological Conflict." *Philosophy & Technology* 34 (December 2021): 717–37. https://doi.org/10.1007/s13347-020-00430-7.

Prunkl, Carina. "Human Autonomy in the Age of Artificial Intelligence." *Nature Machine Intelligence* 4 (February 2022): 99–101. https://doi.org/10.1038/s42256-022-00449-9.

Rachovitsa, Adamantia, and Niclas Johann. "The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case." *Human Rights Law Review* 22, no. 2 (2022): ngac010. https://doi.org/10.1093/hrlr/ngac010.

Robertson, Samantha, and Niloufar Salehi. "What If I Don't Like any of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design." arXiv, July 13, 2020. https://doi.org/10.48550/arXiv.2007.06718.

Rosner, Daniela K., Saba Kawas, Wenqi Li, Nicole Tilly, and Yi-Chen Sung. "Out of Time, Out of Place: Reflections on Design Workshops as a Research Method." In *CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported*

Cooperative Work & Social Computing, 1131–41. New York: ACM, 2016. https://doi.org/10.1145/2818048.2820021.

Rubel, Alan, Clinton Castro, and Adam K. Pham. *Algorithms and Autonomy: The Ethics of Automated Decision Systems*. Cambridge: Cambridge University Press, 2021. https://doi.org/10.1017/9781108895057.

Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1, no. 5 (2019): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Sadowski, Jathan, and Frank Pasquale. "The Spectrum of Control: A Social Theory of the Smart City." *First Monday* 20, no. 7 (2015): 1–22. https://doi.org/10.5210/fm.v20i7.5903.

Sarra, Claudio. "Put Dialectics into the Machine: Protection against Automatic-Decision-Making through a Deeper Understanding of Contestability by Design." *Global Jurist* 20, no. 3 (2020): 20200003. https://doi.org/10.1515/gj-2020-0003.

Sawhney, Nitin. "Contestations in Urban Mobility: Rights, Risks, and Responsibilities for Urban AI." *AI & SOCIETY* 38 (June 2023): 1083–98. https://doi.org/10.1007/s00146-022-01502-2.

Saxena, Devansh, and Shion Guha. "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*, 383–88. New York: ACM, 2020. https://doi.org/10.1145/3406865.3418331.

Saxena, Devansh, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. "A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (2021): article no. 348. https://doi.org/10.1145/3476089.

Schön, Donald A. "Generative Metaphor: A Perspective on Problem-Setting in Social Policy." In *Metaphor and Thought*, 2nd ed., edited by Andrew Ortony, 137–63. Cambridge: Cambridge University Press, 1993. https://doi.org/10.1017/CBO9781139173865.011.

Scott, Deborah. "Diversifying the Deliberative Turn: Toward an Agonistic RRI." *Science, Technology, & Human Values* 48, no. 2 (2023): 295–318. https://doi.org/10.1177/01622439211067268.

Shaw, Joe, and Mark Graham. "An Informational Right to the City? Code, Content, Control, and the Urbanization of Information." *Antipode* 49, no. 4 (2017): 907–27. https://doi.org/10.1111/anti.12312.

Shelton, Taylor, Matthew Zook, and Alan Wiig. "The 'Actually Existing Smart City.'" *Cambridge Journal of Regions, Economy and Society* 8, no. 1 (2015): 13–25. https://doi.org/10.1093/cjres/rsu026.

Soja, Edward. "The City and Spatial Justice." *Justice Spatiale/Spatial Justice* 1, no. 1 (2009): 1–5. https://www.jssj.org/article/la-ville-et-la-justice-spatiale/?lang=en.

Stilgoe, Jack, Richard Owen, and Phil Macnaghten. "Developing a Framework for Responsible Innovation." *Research Policy* 42, no. 9 (2013): 1568–80. https://doi.org/10.1016/j.respol.2013.05.008.

Stolterman, Erik, and Mikael Wiberg. "Concept-Driven Interaction Design Research." *Human-Computer Interaction* 25, no. 2 (2010): 95–118. https://doi.org/10.1080/07370020903586696.

Suchman, Lucy. "Corporate Accountability." *Robot Futures*, June 10, 2018. https://robotfutures.wordpress.com/2018/06/10/corporate-accountability/.

Thackara, John. *In the Bubble: Designing in a Complex World*. Cambridge, MA: MIT Press, 2005.

Thoring, Katja, Roland Mueller, and Petra Badke-Schaub. "Workshops as a Research Method: Guidelines for Designing and Evaluating Artifacts through Workshops." In *Proceedings of the 53rd Hawaii International Conference on System Sciences 2020*, 5036–45. Hawaii: HICSS, 2020. https://hdl.handle.net/10125/64362.

Tufte, Edward R. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press, 1997.

Umbrello, Steven. "Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design." *Science and Engineering Ethics* 26 (April 2020): 575–95. https://doi.org/10.1007/s11948-019-00104-4.

Vaccaro, Kristen, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. "Contestability in Algorithmic Systems." In *CSCW '19 Companion: Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 523–27. New York: ACM, 2019. https://doi.org/10.1145/3311957.3359435.

Van Bekkum, Marvin, and Frederik Zuiderveen Borgesius. "Digital Welfare Fraud Detection and the Dutch SyRI Judgment." *European Journal of Social Security* 23, no. 4 (2021): 323–40. https://doi.org/10.1177/13882627211031257.

Van Bouwel, Jeroen, and Michiel van Oudheusden. "Participation beyond Consensus? Technology Assessments, Consensus Conferences and Democratic Modulation." *Social Epistemology* 31, no. 6 (2017): 497–513. https://doi.org/10.1080/02691728.2017.1352624.

Veale, Michael, Max van Kleek, and Reuben Binns. "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*, paper no. 440. New York: ACM, 2018. https://doi.org/10.1145/3173574.3174014.

Verma, Himanshu, Jakub Mlynar, Roger Schaer, Julien Reichenbach, Mario Jreige, John Prior, Florian Evéquoz, and Adrien Depeursinge. "Rethinking the Role of AI with Physicians in Oncology: Revealing Perspectives from Clinical and Research Workflows." In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, article no. 17. New York: ACM, 2023. https://doi.org/10.1145/3544548.3581506.

Walmsley, Joel. "Artificial Intelligence and the Value of Transparency." *AI & SOCIETY* 36, no. 2 (2021): 585–95. https://doi.org/10.1007/s00146-020-01066-z.

Williams, Jennifer, Dena Fam, and Abby Mellick Lopes. "Creating Knowledge: Visual Communication Design Research in Transdisciplinary Projects." In *Transdisciplinary Research and Practice for Sustainability Outcomes*, edited by Dena Fam, Jane Palmer, Chris Riedy, and Cynthia Mitchell, chapter 11. London: Routledge, 2016. https://doi.org/10.4324/9781315652184.

Wiltshire, Gareth, and Noora Ronkainen. "A Realist Approach to Thematic Analysis: Making Sense of Qualitative Data Through Experiential, Inferential and Dispositional Themes." *Journal of Critical Realism* 20, no. 2 (2021): 159–80. https://doi.org/10.1080/14767430.2021.1894909.

Yildirim, Nur, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. "Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook." In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, article no. 356. New York: ACM, 2023. https://doi.org/10.1145/3544548.3580900.

Young, Matthew M., Justin B. Bullock, and Jesse D. Lecy. "Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration." *Perspectives on Public Management and Governance* 2, no. 4 (2019): 301–13. https://doi.org/10.1093/ppmgov/gvz014.

Yurrita, Mireia, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. "Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability." In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, article no. 134. New York: ACM, 2023. https://doi.org/10.1145/3544548.3581161.

Appendix A

Infographic Description

The infographic in Figure 3 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 A1). We are taking a socio-technical view.[i] The *system* consists not only of technology but also of 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 controller 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.

Now, we move on to the contestability mechanisms. First, *interactive controls* (Figure A2) 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.

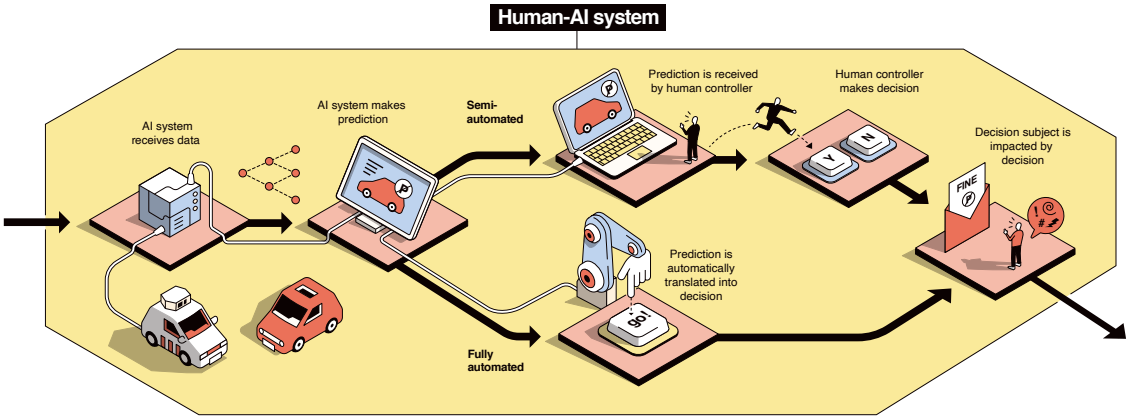
Next, we look at contestation after a decision has been made. So-called *intervention requests* (Figure A3). These can be broken down into explanations, channels for voice, arenas for debate, and the obligation to respond. First, the system must provide a subject with an explanation of how a decision 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.

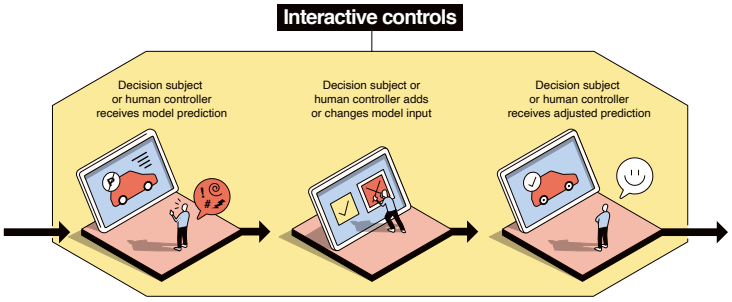
Connected to the previous decision-appeal loop is a second-order *monitoring* loop (Figure A4). Here, a record of all decision appeals is kept. This record is analyzed for patterns that indicate systemic shortcomings. A human operator is alerted to investigate if such a pattern is suspected. It is then up to the operator to decide on further action. A systemic flaw can require technology revision or, further upstream, to revise policy.

The following mechanism is about *global* contestability. *Tools for scrutiny* (Figure A5) 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.

Since we are explicitly dealing with public AI systems in this infographic, we also have an area for *policy & system development* (Figure A6). In this area, 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. Processes in this area produce the policies that directly govern human controller behavior or are translated into technology.

i Lucy Suchman et al., "Reconstructing Technologies as Social Practice," *American Behavioral Scientist* 43, no. 3 (1999): 392–408, https://doi.org/10.1177/00027649921955335.





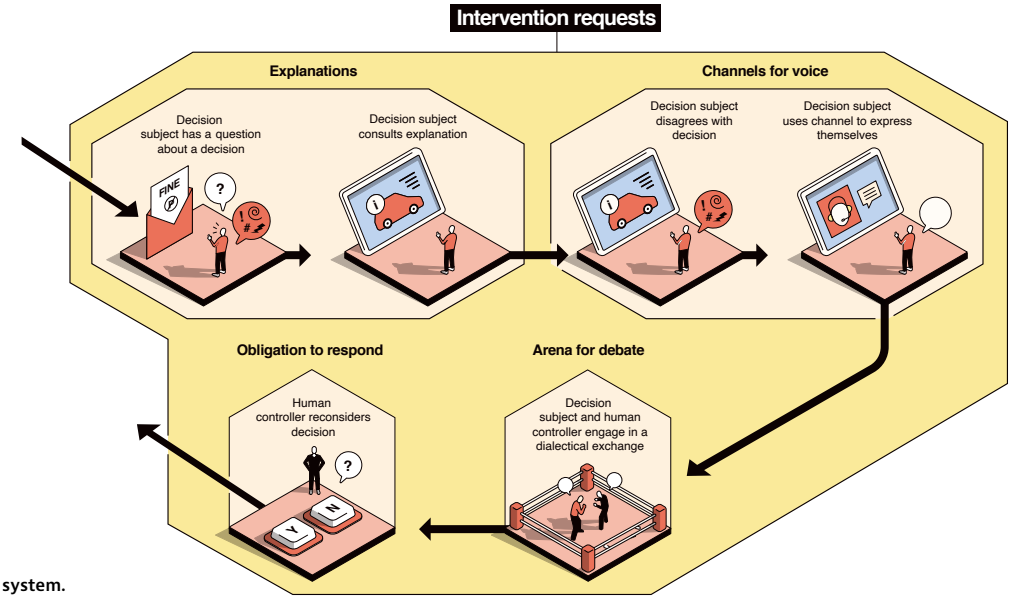


Figure A1
Infographic detail: human-AI system.

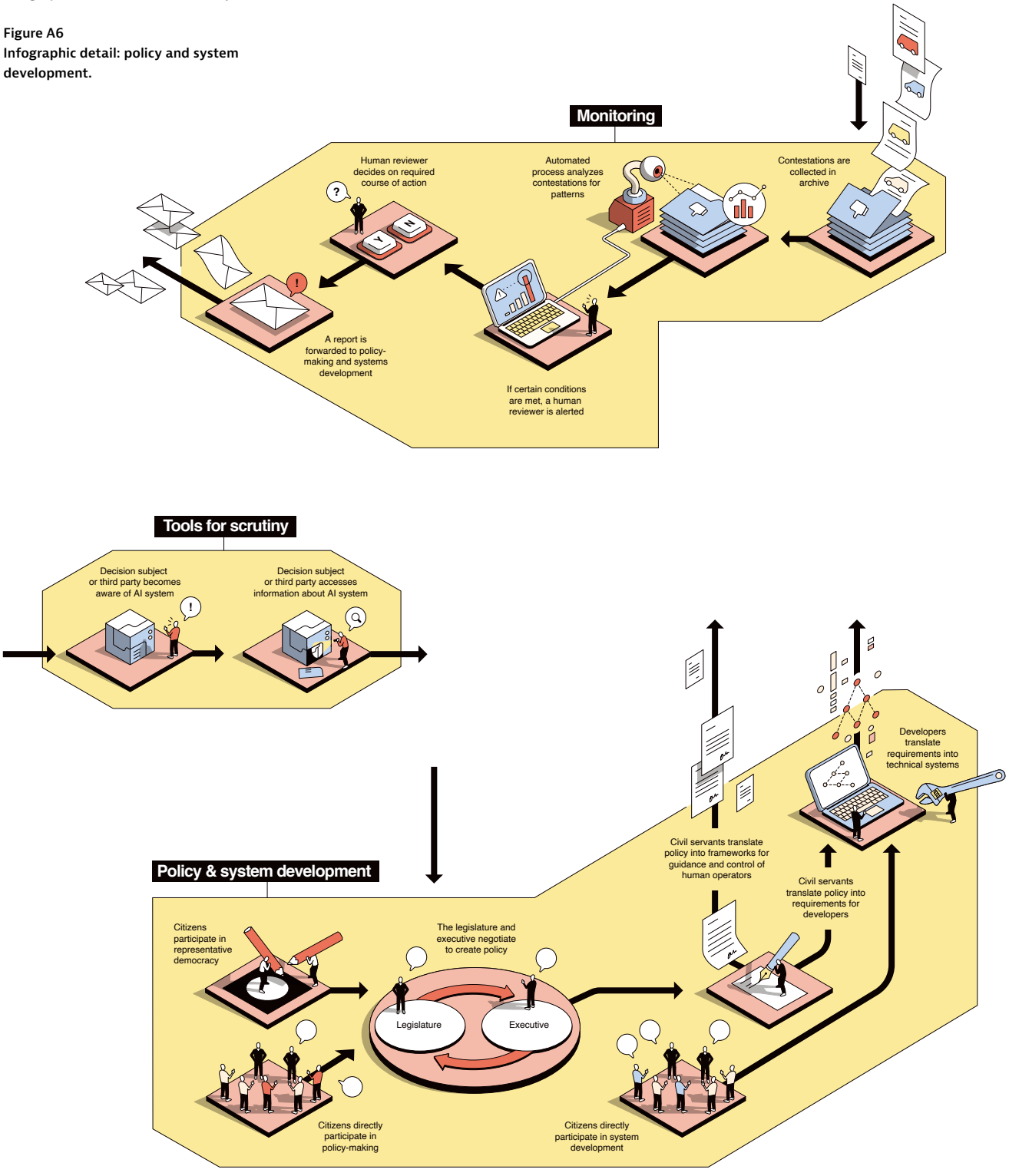
Figure A2
Infographic detail: interactive controls.

Figure A3
Infographic detail: intervention requests.

Figure A4
Infographic detail: monitoring.

Figure A5
Infographic detail: tools for scrutiny.

Figure A6
Infographic detail: policy and system development.



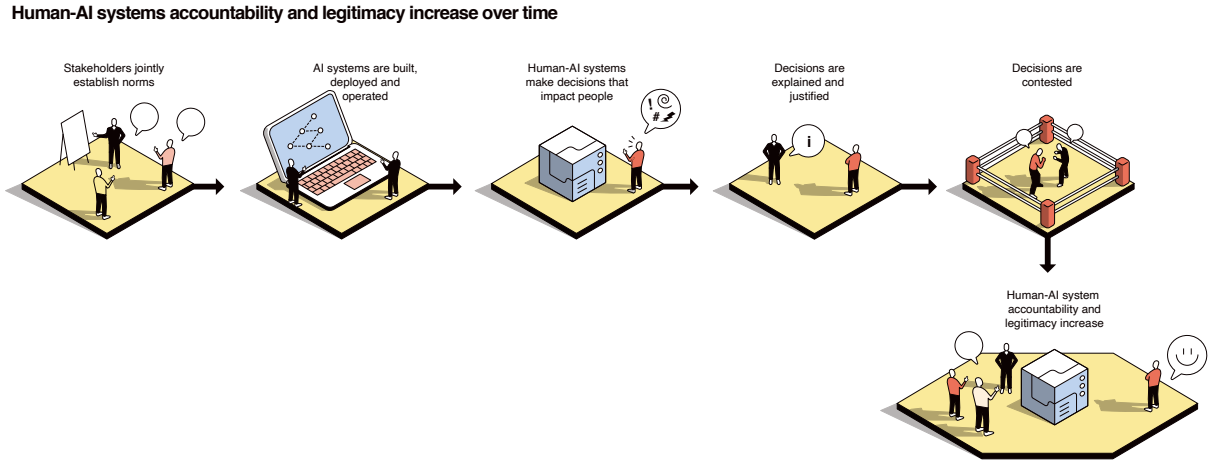


Figure A7
Infographic detail: accountability and legitimacy increase over time.

The flow at the bottom (Figure A7) 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.