The Democracy Levels Framework: Towards Integrating Democratic Infrastructure into the AI Ecosystem

Aviv Ovadya[1] Kyle Redman[1, 2] Oliver Smith[1] Luke Thorburn[1,3]
Flynn Devine[4] Andrew Konya[5] Smitha Milli[6] Manon Revel[6]
Quan Ze Chen[1, 7] Kevin Feng[7] Amy X. Zhang[7]
Bilva Chandra Michiel A. Bakker[8] Atoosa Kasirzadeh[9]

[1] AI & Democracy Foundation
[2] newDemocracy Foundation
[3] King's College London
[4] Boundary Object Studio
[5] Remesh
[6] Meta AI, FAIR Labs
[7] University of Washington
[8] MIT
[9] Carnegie Mellon University

aviv@aidemocracyfoundation.org

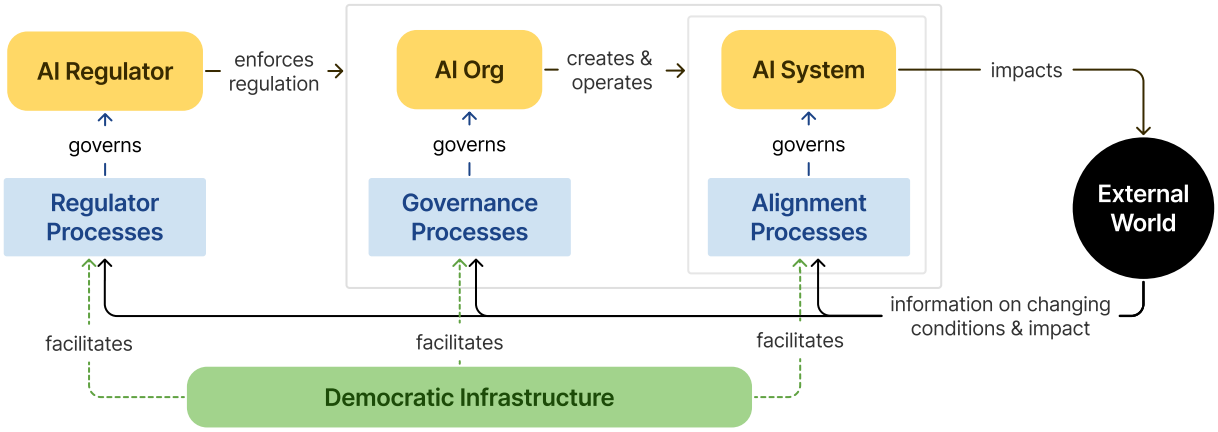

Figure 1: A system diagram of how democratic processes could integrate with the AI ecosystem, with democratic infrastructure being used to facilitate — where appropriate — collective decisions relating to AI regulation, organizational governance, and alignment. The Democracy Levels Framework we introduce can be used to evaluate (i) the degree to which democratic systems are used for decision-making, and (ii) the quality of those democratic systems, and the infrastructure supporting them.

Abstract

Decisions around AI come with increasingly systemic and societal impacts. However, effectively "democratizing AI" requires integrating democratic infrastructure into multiple aspects around the governance and alignment of AI. While initial steps — such as Meta's *Community Forums* and Anthropic's *Collective Constitutional AI* — have illustrated a promising direction, where democratic processes might be used to meaningfully improve public involvement and trust in critical decisions, a more concrete roadmap for increasingly democratic AI remains to be defined. In this work, we introduce the "Democracy Levels Framework", a set of tools that: (i) defines milestones toward meaningfully democratic, pluralistic, human-centered, and public-interest AI, (ii) helps guide organizations seeking to increase the legitimacy of their decisions on difficult AI governance and alignment questions, and (iii) supports the evaluation of such efforts.

Keywords

democracy, framework, AI, alignment

CHI STAIG '25, Yokohama, Japan

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Machine learning research is advancing at a breathtaking pace, and its results are having increasingly dramatic societal impacts at an unprecedented speed. AI systems make explicit and implicit decisions every day that have impacts on billions of people. But who should steer the development of AI? Similar questions have emerged with previous technological advances [13, 28, 42], and existing institutions and power structures will clearly play a significant role in adjudicating these questions. However, with AI, the pace of change, ubiquity, market incentives, geopolitical incentives, and jurisdictional arbitrage opportunities pose unprecedented challenges [1]. Thankfully, recent innovations in collective decision-making point towards a new generation of processes, infrastructures, and institutions to navigate these challenges [9, 22, 26, 38]. They provide new ways to approach ensuring that the development of AI remains human-centered, not *just* at an individual human level, but societally and even globally.

In this work, we propose the "Democracy Levels" framework, a set of tools which can: (a) be used as **milestones** toward a roadmap for the democratic AI [9], pluralistic AI [37], and public AI [27, 40] ecosystems; (b) help **guide organizations** and institutions that need to increase the legitimacy of difficult AI governance and alignment decisions; and (c) be used as an **evaluation framework** to identify opportunities for improvement and keep AI organizations accountable. Our framework applies to multiple facets of the AI ecosystem (Figure 1), with the ultimate goal of providing a clear map of what it would take to enable meaningful democratic governance and alignment of AI, in a way that is useful both internally to organizations making decisions about AI, and externally to those supporting this work and providing accountability.

2 Related Work

AI Alignment. Alignment mechanisms such as RLHF [24], or DPO [29], or Constitutional AI [4], aim to make AI systems more aligned with humans. However, democratic considerations have often not been central concerns, with unrepresentative selection and simple aggregation being used [35]. More recently, efforts are growing to make alignment more democratic, such as Anthropic's Collective Constitutional AI [16] which uses Pol.is to identify principles from a representative sample of U.S. adults, or Conitzer et al. [10] which calls for explicitly applying social choice approaches to preference learning methods [14] instead of aggregating annotators as one input. While nascent efforts are exciting, there is as of yet no standard way to evaluate how democratic these new alignment approaches are.

Human-Centered and Participatory AI. Participatory AI encompasses a range of processes to engage people in decision-making around AI, from approaches that *consult* with stakeholders (e.g., expressing preferences for policies in a ranking), to those in which stakeholders *own* the design process and play a central role [12]. However, progress is nascent, with private domain efforts largely confined to consultation [15] and public efforts facing difficult integration challenges around who should participate [41]. Some have desired further clarity around the definition, role, and broader relationship of Participatory AI [5] which our framework seeks to provide. Our framework also strives to make AI systems more human-centered—ensuring that the AI can better serve human needs [3, 7, 30, 32, 33]. Democratic AI and "traditional" human-centered AI share many goals; advancing one can often advance the other [34].

Frameworks around Democracy and Responsible AI. To develop this framework, we have also drawn inspiration from existing frameworks for evaluating democratic-ness [2, 17, 20, 36], as well as frameworks for evaluating degrees of responsible behavior and autonomy in AI systems [6, 31]. Our work relates to explorations and assessments of democratic [9], participatory [11, 12, 39], pluralistic [37], human-centered [34], and public AI [27, 40].

3 The Democracy Levels Framework

The Democracy Levels Framework is made up of (i) a set of Levels which capture the degree to which decision-making power has been transferred to a democratic system, and (ii) a set of dimensions which capture the qualities of that democratic system. We then introduce two tools to aid the integration of democratic infrastructure into the AI ecosystem: a **Democratic System Card** that can be used to evaluate a democratic system against the dimensions, and a **Levels Decision Tool** for informing decisions about "how much democracy" is appropriate in a given decision-making context.

Levels. The transfer of decision-making power from a *unilateral authority* to *democratic systems* can take many forms, but there are discrete points of particular significance which may require new kinds of democratic infrastructure. We have developed **levels** (Figure 2) to provide clearer distinctions for understanding and implementing these transfers, building on experience supporting movement between these levels. We define each level of democratic decision-making according to which of five roles are performed by democratic systems, rather than a unilateral authority: (i) **informing** decisions; (ii) **specifying** options; (iii) making **binding** decisions; (iv) automatically triggered **initiation** of binding decision-making processes; and (v) **constitutional** decision-making. Figure 2 provides definitions for each level, from L0 to L5, along with concrete examples of what this could look like in practice for a plausible decision domain: developing a set of rules governing persuasion by an AI system. Such rules might be used directly in model training (e.g., to align an AI system) [21] or as policies (e.g., for an AI organization or regulator).

The resulting framework is somewhat analogous to the autonomy levels defined for self-driving cars [31], as it also involves the shifting of decision-making from a unilateral authority (i.e., human driver or AI corporation) to a new kind of decision-making system (i.e., autonomous control system or democratic system).

Different domains might involve democratic systems at different levels. For example, decisions about whether to release a new model might be at L2, while decisions about the model spec used for fine-tuning could be at L4. Scope limitations can (and often should) also be provided, for example, specifying that any rules developed via a process would only be binding for two years, or until a given condition is met (such as a model passing a particular benchmark).

Dimensions. In order for more decision-making power to be transferred to democratic systems (thus a rise in level), they must have sufficiently desirable properties to support the meaningful,

	Roles Performed by Democratic Processes	Description	Example
L0	∅	**Unilateral** decision-making: all formal decision-making authority lies with the unilateral authority.	*Rules on AI persuasion are simply created by the unilateral authority.*
L1		Outputs of a democratic process **inform** the unilateral authority; such democratic processes are initiated ad-hoc when desired and with a remit chosen by the unilateral authority.	*The process outputs recommendations on AI persuasion, which need to be interpreted by the unilateral authority for implementation as rules.*
L2		Democratic processes output a fully **specified** decision which must be implemented by default unless the unilateral authority uses a predetermined process or criteria to amend or veto.	*The process outputs rules on AI persuasion, which are implemented as-is, unless amended or vetoed.*
L3		Democratic process outputs are **binding** and cannot be vetoed (assuming feasibility, e.g. technically, legally; and within their remit).	*The process outputs rules on AI persuasion, which are implemented as-is (unless a pre-established process finds it infeasible).*
L4		The unilateral authority pre-commits to the automatic **initiation** of binding democratic processes when a given condition is met (instead of being initiated ad-hoc), with scope over a pre-specified domain.	*Processes to update rules on AI persuasion are run yearly or whenever a newly pretrained model is to be deployed.*
L5		The unilateral authority fully shifts power within a domain of decision-making to an adaptive "**constitutional** order" — a system of checks and balances which is used to determine when and how democratic processes are to be used (within a pre-specified domain).	*The decisions around when to trigger processes to update rules (and how those processes are triggered) are also under the control of democratic processes (via a system of checks and balances).*

informing decisions · specifying options · binding decisions · automatic initiation · constitutional structures

Figure 2: Overview of the Democracy Levels, which are used to assess how much decision-making power in a given domain of decision-making has been transferred from a unilateral authority to a democratic process.

safe, and effective implementation of such higher democracy levels. So in addition to *levels* which define the role of democratic systems in relation to unilateral authorities, we also define three primary *dimensions* (Appendix D)—process quality, delegation, and trust—to capture different facets of properties that democratic systems may have. To successfully move up a democracy level, democratic systems must first be able to reliably provide a certain level of **process quality**, such that it is appropriate to shift power—i.e., they must ensure that processes are *representative*, people are *informed*, reasoning is *deliberative*, outputs are *substantive*, and decisions *robust* to adversarial behavior. Then, the unilateral authority must increase its capacity to **delegate** to a democratic process. This includes its capacity to organizationally and publicly *commit* to the outcomes; to *integrate* such processes into its operations; and to technically and/or legally *bind* itself to the resulting decisions. Finally, there must be external conditions that support the process, which we collectively refer to as **trust**—i.e., the relevant public and stakeholders must be sufficiently *aware* of the process, buy into its *legitimacy*, and be willing to *participate*; and there must be sufficiently capable forms of *accountability*.

3.1 Applying the Framework

Balancing the levels and properties of the democratic system as outlined above can be challenging, so we additionally provide two artifacts intended to help users of our framework navigate the design space:

The Democratic System Card. The system card (Appendix C) is intended to help decision-makers document, assess, compare, and evaluate democratic systems in a structured manner, with a core goal of providing insight into how appropriate a system is at a given democracy level, for a given context. A full system card has three primary components that are assessed over each dimension: (1) **descriptions** of how the democratic system operates; (2) **assessments** of the system implementation, based on *guiding questions*; finally, (3) a qualitative **evaluation** of the highest level of power said system should be trusted (given some context). Decision-makers can use system cards to assess whether a democratic system is ready to be delegated more binding power in decision-making. Stakeholders and advocacy groups can make use of system cards to compare and contrast possible alternative systems to propose or

advocate for. While active and empowered democratic systems, or proposals for complete democratic systems, should have complete cards, prototypes and research projects may have only parts of the card filled out, as not every aspect is applicable. System cards can also be a resource for those identifying gaps in the democratic infrastructure ecosystem, aiding understanding of critical needs for operating at higher levels of delegated power, and consideration of how democratic processes complement one another.

The Levels Decision Tool. The levels tool (Appendix D) helps a *unilateral authority* understand under which conditions it may want to delegate decision-making power, e.g., to increase public trust. But how does it go about choosing a level? This often depends on reflecting over the **context** of the decision, including: the unilateral authority; its scope of authority; who is affected by the decision; how the authority relates to other stakeholders, both internal and external; etc. The levels decision tool provides targeted questions around inspecting this context that can help determine which democracy level is appropriate, such as how to consider legitimacy, stakeholders, as well as speed and cost considerations. The questions involve: the value of legitimacy across the public, internal stakeholders, external stakeholders, and government; the potential benefits of collective intelligence; the feasibility of transferring decision-making power; the importance of speed and adaptability to the decision domain being considered; resourcing; and novelty. Some questions are applicable to every context, and some are only applicable to corporations or to governments.

4 Discussion and Future Work

It is important to keep in mind that integrating and evaluating democratic systems is ultimately context-dependent, and human stakeholders must consider the questions in our framework for themselves. Our goal is not to prescribe when a higher democracy level is warranted or whether a particular democratic system is better, but rather to present a common framework on which discussions around the democratization of AI can be organized. For example, helping clarify when organizations are claiming to be acting more democratically than they actually are.

Finally, we note that maturity also doesn't come overnight—organizations, democratic infrastructure providers, stakeholders, and the public all need to build democratic muscle—and taking on too much all at once can backfire. Instead of holding organizations to a platonic ideal, it can often be more helpful to focus on improvements at the margin (relative to the current democracy level or status quo alternatives), both of which can be articulated through a level system. By providing a concrete articulation that we expect to be further built upon, we hope that we may enable more productive conversations about what future we should be aiming for with regards to power, participation, pluralism, and democracy. Democracy is a journey, and we aim to have provided a useful map.

References

[1] Danielle Allen and E Glen Weyl. 2024. The Real Dangers of Generative AI. *Journal of Democracy* 35, 1 (2024), 147–162.

[2] Sherry R. Arnstein. 1969. A Ladder Of Citizen Participation. *Journal of the American Institute of Planners* 35, 4 (1969), 216–224. doi:10.1080/01944366908977225 arXiv:https://doi.org/10.1080/01944366908977225

[3] Jan Auernhammer. 2020. Human-centered AI: The role of Human-centered Design Research in the development of AI. (2020).

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[5] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. doi:10.1145/3551624.3555290

[6] Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. 2024. The Foundation Model Transparency Index v1.1: May 2024. doi:10.48550/arXiv.2407.12929 arXiv:2407.12929 [cs] Accessed: 2024-09-04.

[7] Tara Capel and Margot Brereton. 2023. 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*. 1–23.

[8] Tom Christiano and Sameer Bajaj. 2024. Democracy. In *The Stanford Encyclopedia of Philosophy* (summer 2024 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2024/entries/democracy/ Accessed: 2024-09-03.

[9] CIP. 2024. *A Roadmap to Democratic AI.* Technical Report. The Collective Intelligence Project. https://cip.org/s/CIP_-A-Roadmap-to-Democratic-AI.pdf

[10] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. 2024. Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. *arXiv preprint arXiv:2404.10271* (2024).

[11] Ned Cooper and Alex Zafiroglu. 2024. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–9. doi:10.1145/3613904.3642775 arXiv:2403.06431 [cs] Accessed: 2024-09-04.

[12] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. doi:10.1145/3617694.3623261

[13] Laura DeNardis. 2014. *The Global War for Internet Governance.* Yale University Press.

[14] Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. 2024. Axioms for AI Alignment from Human Feedback. *arXiv preprint arXiv:2405.14758* (2024).

[15] Lara Groves, Aidan Peppin, Andrew Strait, and Jenny Brennan. 2023. Going public: the role of public participation approaches in commercial AI labs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1162–1173.

[16] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1395–1417. doi:10.1145/3630106.3658979

[17] IAP2 Australasia. 2024. IAP2 Public Participation Spectrum. https://iap2.org.au/resources/spectrum/ Accessed: 2024-09-04.

[18] Andrew Konya, Aviv Ovadya, Kevin Feng, Quan Ze Chen, Lisa Schirch, Colin Irwin, and Amy X Zhang. 2024. Chain of Alignment: Integrating Public Will with Expert Intelligence for Language Model Alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*. https://openreview.net/forum?id=QOSaR4Ur0v

[19] Andrew Konya, Deger Turan, Aviv Ovadya, Lina Qui, Daanish Masood, Flynn Devine, Lisa Schirch, Isabella Roberts, and Deliberative Alignment Forum. 2023. Deliberative Technology for Alignment. arXiv:2312.03893 [cs.CY] https://arxiv.org/abs/2312.03893

[20] Staffan I. Lindberg, Michael Coppedge, John Gerring, and Jan Teorell. 2014. V-Dem: A New Way to Measure Democracy. *Journal of Democracy* 25, 3 (2014), 159–169. doi:10.1353/jod.2014.0040 Accessed: 2024-09-04.

[21] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Improving Model Safety Behavior with Rule-Based Rewards. *OpenAI Blog* (2024). https://openai.com/index/improving-model-safety-behavior-with-rule-based-rewards/ Preprint.

[22] OECD. 2020. *Innovative Citizen Participation and New Democratic Institutions.* OECD iLibrary. 196 pages. doi:https://doi.org/10.1787/339306da-en

[23] OpenAI. 2024. Model Spec. https://cdn.openai.com/spec/model-spec-2024-05-08.html Accessed: 2025-01-31.

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[25] Aviv Ovadya. 2023. 'Generative CI'through Collective Response Systems. *arXiv preprint arXiv:2302.00672* (2023).

[26] Aviv Ovadya. 2023. Reimagining Democracy for AI. *Journal of Democracy* 34, 4 (2023), 162–170. doi:10.1353/jod.2023.a907697 Accessed: 2024-09-04.

[27] Public AI Network. 2024. *Public AI: Infrastructure for the Common Good.* Technical Report. The Public AI Network. https://publicai.network/whitepaper

[28] Roxana Radu. 2019. *Negotiating Internet Governance.* Oxford University Press.

[29] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).

[30] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human behavior and emerging technologies* 1, 1 (2019), 33–36.

[31] SAE International. 2021. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104.* Technical Report. SAE International. https://www.sae.org/standards/content/j3016_202104/ Standards Specification.

[32] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.

[33] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.

[34] Anton Sigfrids, Jaana Leikas, Henrikki Salo-Pöntinen, and Emmi Koskimies. 2023. Human-centricity in AI governance: A systemic approach. *Frontiers in Artificial Intelligence* 6 (2023), 976887.

[35] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *ICLR.*

[36] Svend-Erik Skaaning and Alexand Hudson. 2023. *The Global State of Democracy Indices Methodology: Conceptualization and Measurement Framework, Version 7 (2023)* (7 ed.). International Institute for Democracy and Electoral Assistance (International IDEA). doi:10.31752/idea.2023.38 Accessed: 2024-09-04.

[37] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. doi:10.48550/arXiv.2402.05070 arXiv:2402.05070 [cs] Accessed: 2024-09-04.

[38] Jack Stilgoe. 2024. AI has a democracy problem. Citizens' assemblies can help. *Science* 385, 6711 (2024), eadr6713. doi:10.1126/science.adr6713 arXiv:https://www.science.org/doi/pdf/10.1126/science.adr6713

[39] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24).* Association for Computing Machinery, Rio de Janeiro, Brazil, 1609–1621. doi:10.1145/3630106.3658992

[40] Nicholas Vincent, David Bau, Sarah Schwettmann, and Joshua Tan. 2023. An Alternative to Regulation: The Case for Public AI. In *Proceedings of the NeurIPS 2023 Workshop on Regulatable ML.* https://openreview.net/forum?id=TFWnViI30j Accessed: 2024-09-04.

[41] Janis Wong, Deborah Morgan, Vincent J Straub, Youmna Hashem, and Jonathan Bright. 2022. Key challenges for the participatory governance of AI in public administration. *SocArXiv Papers: https://osf. io/preprints/socarxiv/pdcrm* (2022).

[42] Malte Ziewitz and Ian Brown. 2013. *A prehistory of internet governance.* Edward Elgar Publishing, Cheltenham, UK, Chapter 1, 3–26. doi:10.4337/9781849805025.00008

A Terminology

Scope of Authority. : The set of *powers* that an authority is granted, including a specification of the *domain(s)* governed by that authority, and potentially implicit *external constraints*. For example, the scope of authority of a finance committee for a US corporation's board of directors is set out by a corporate charter, bylaws, and board resolutions; constrained by regulations and case law; and can have its authority limited to financial matters.

Unilateral authority. : An entity that can make decisions without meaningful checks on its power within a given scope of authority. The unilateral authority over a given scope does not need approval for any decisions within that scope.

Democratic Process. : A collective decision-making process characterized by "a kind of equality among the participants at an essential stage of the decision-making process" [8].

Democratic System. : A set of interacting entities and democratic processes.

Under these definitions, elections, citizen assemblies [22], and collective dialogue processes [19, 25] are *democratic processes*. The interactions between those processes and *unilateral authorities*, constituents, stakeholders, media, etc. make up a *democratic system* (supported via *democratic infrastructure*). A more complex democratic system might involve an entire constitutional order with multiple institutions interacting through different processes on an ongoing basis. Concretely, for AI, an ongoing Chain of Alignment process [18], feeding into a model spec (e.g., OpenAI [23]), managed by a democratic oversight body, all coordinated by democratic infrastructure providers, could have a *scope of authority* over AI model alignment.

B Framework Application Illustration

In Figure 3, we illustrate how different stakeholders may apply the tools in our framework. For example, we envision that the Levels Decision Tool may be used to answer questions like "What level to aim for given some context (Figure 1)?" With the Democratic System Card, we envision that users may ask questions like "Is a particular democratic system good enough to be used at a certain level given some context?"

C Democratic System Card

The Democratic System Card (Figure 4 and Figure 5) is a set of questions to guide reflection on whether the quality of a democratic system is commensurate with the level of decision-making power delegated to it, for a given domain of decision-making, in a given context. The questions are grouped by the dimensions in the Democracy Levels Framework. To complete a System Card, you first (1) describe the process or system at a high level, and (2) summarize what other systems the process depends on or interacts with, which impact its success (e.g. sortition data, or user or citizen authentication systems). You then complete the following table.

D Levels Decision Tool

The Levels Decision Tool (Figure 6 and Figure 7) is a set of questions to help determine which Democracy Level to aim for in a given context. The arrows refer to whether each question, if answered in the affirmative, gives reason for targeting a higher or lower Democracy Level. In some cases, both arrows are used to indicate that the implication for what Level to aim for will be context-dependent.

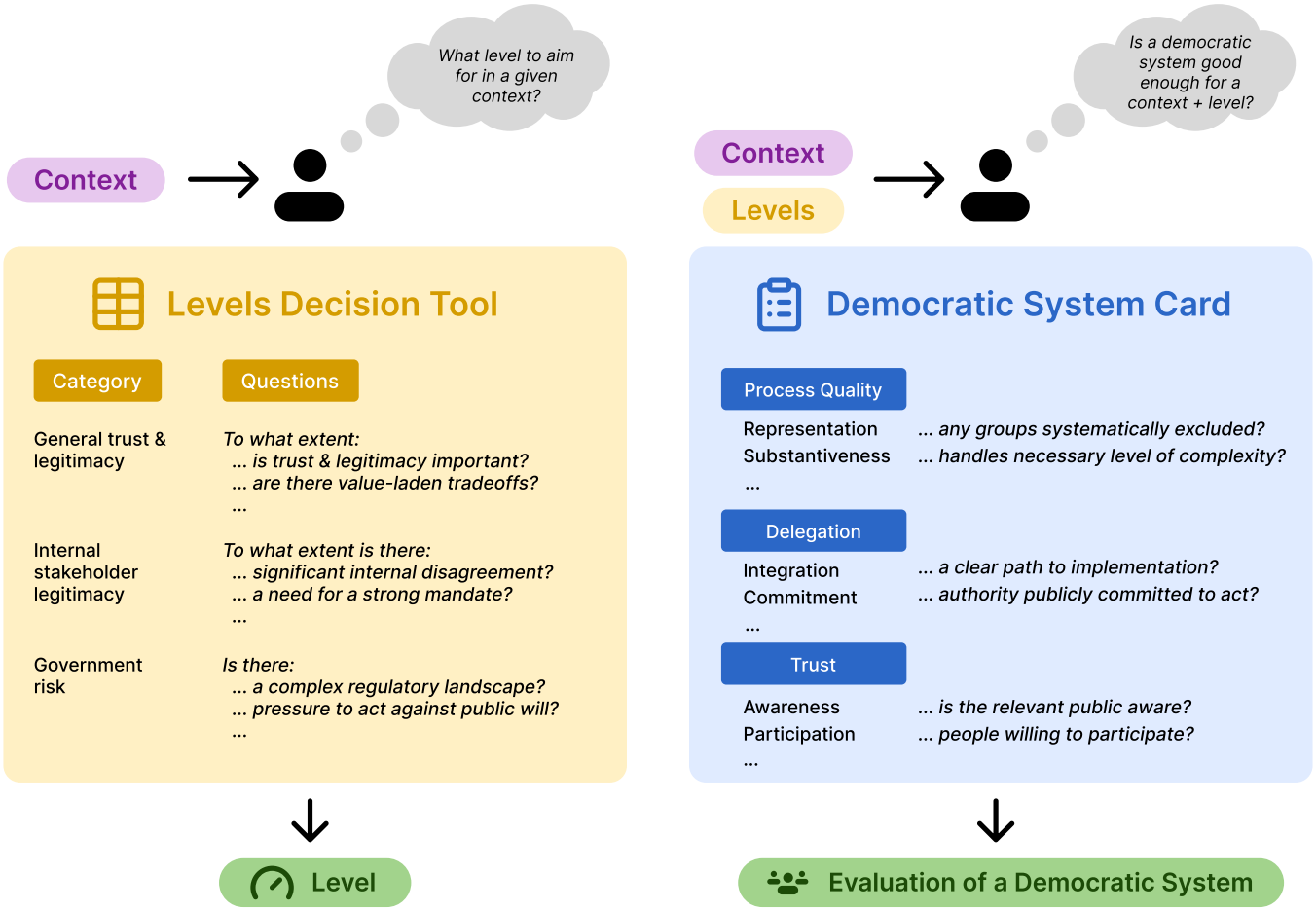

Figure 3: An overview of the Levels Decision Tool and Democratic System Card, and how they can be used.

Process Quality	Definition	Description	Guiding Questions	Assessment
	The extent to which …	**Task:** Below, write a brief description of how this dimension/property works in the democratic process or system you are evaluating.	Aspects of the dimension/property that should be considered in your assessment.	**Task:** Below, using the guiding questions as a guide, reflect on the extent to which each dimension/property is satisfied in the democratic process or system you are evaluating.
Representation	… key decisions (+ those who make them) are sufficiently representative of the constituent population.		To what extent: 1. are there barriers leading groups to be significantly under-represented? 2. is there sufficient representation at critical parts of the process, including (a) proposing alternative approaches, and (b) making ultimate decisions?	
Informedness	… those making decisions understand the information critical to making that decision.		To what extent do participants gain critical context about tradeoffs and implications of different decisions from (a) experts, (b) the existing authorities, who may have extensive context, (c) a broad diversity of constituents, (d) the most impacted stakeholders, and (e) the powerful stakeholders, whose incentives are critical to having the decision "stick"?	
Deliberation	… decisions are considered and deliberative (rather than unconsidered and reactive).		To what extent are those involved: 1. able to (and supported to) move from shallower to deeper goals and values? 2. able to collaborate, and are sufficiently supported to do this?	
Substantiveness	… decisions are substantive (e.g., actionable + consequential) rather than nonsubstantive (e.g., vague, simplistic, or inconsequential).		To what extent: 1. is the decision directly actionable, with little need or room for interpretation? 2. does the decision grapple with the necessary level of complexity to tackle the issue? 3. is uncertainty appropriately managed and accounted for?	
Robustness	… the process is robust to suboptimal conditions or adversarial or strategic behavior.		To what extent is the process or system vulnerable to: 1. suboptimal conditions or broken assumptions, such as low turnout or the presence of interpersonal power asymmetries? 2. strategic behavior and manipulation? 3. false claims of manipulation?	
Legibility	… the decision and process are accessible, understandable, and verifiable to those not directly involved.		To what extent can non-participants understand: 1. what decision was taken? 2. why the decision was taken? 3. which groups were 'for' and which 'against' the decision?	

Figure 4: Full version of the democratic system card (First half)

Delegation				
Integration	… the unilateral authority integrates the democratic process into its operations.		To what extent: 1. can the process be automatically triggered, or its decisions be automatically implemented? 2. are the decisions in a format that can be implemented without interpretation or editorialization?	
Ability to bind	… the unilateral authority is able to technically and legally bind itself to act in accordance with the democratic decision.		To what extent: 1. is it *technically* feasible for the unilateral authority to bind itself to acting in accordance with the democratic decision? (e.g., it is difficult to have hard guarantees on language model outputs) 2. is it *legally* feasible for the unilateral authority to bind itself to act in accordance with the democratic decision? (e.g., are the appropriate legal structures in place)	
Commitment	… the unilateral authority commits to acting in accordance with the democratic decision.		To what extent has the unilateral authority internally, privately, and publicly committed to acting in accordance with the democratic decision? (regardless of their ability to bind)	
Trust				
Awareness	… the relevant public is aware of the democratic process.		To what extent is the relevant public aware: 1. that the democratic process or system exists, and how it works? 2. of the decisions that it has made in the past, and/or is in the process of making? 3. of pathways through which they can be involved?	
Participation	… the relevant public is willing to participate in the process.		To what extent is the relevant public: 1. willing to participate? 2. able to participate? 3. fairly compensated for participating? 4. actually participating?	
Accountability	… there are external watchdogs and accountability structures monitoring the execution of the democratic process and the implementation of its outputs.		To what extent are: 1. democratic processes and systems held to a high standard? 2. unilateral authorities held to their promised levels of democratic involvement? 3. both unilateral authorities and democratic processes and systems responsive to such accountability mechanisms?	
Buy-in	… the relevant public and key stakeholders buy-in to the process and its legitimacy.		To what extent do the relevant public and key stakeholders see the process and resulting decisions as: 1. legitimate? 2. preferable to alternative governance mechanisms?	

Figure 5: Full version of the democratic system card (Second half)

Category	Questions for Decision Makers and Advocates
General trust and legitimacy	To what extent: - ↑↑ Is trust and legitimacy important to the UA? - ↑↑ Is the UA not trusted by other actors, regardless of the decisions it makes?x\` To what extent does the decision involve: - ↑↑ Values-laden tradeoffs? - ↑ Significant public interest concerns or externalities? - ↑↓ Limited public impact? - ↑↓ No clear expert consensus? - ↓ A private technical or operational matter?
External stakeholder legitimacy	To what extent is there: - ↑ Powerful stakeholder groups, with conflicting perspectives? - ↑ A bias for inaction given such conflict? - ↑ An opportunity for decreased criticism or consequences due to meaningful “process legitimacy”

Internal stakeholder legitimacy	To what extent is there:
	- ↑ Significant internal disagreement?
	- ↑ A need to evolve organizational values or purpose with a strong mandate?
	- ↑ Talent motivated by public benefit?
	- ↑ Collaborative decision-making history?
	- ↑↓ Hierarchical culture
	    - ↑ Can make it easier to delegate power to a democratic system
	    - ↓ Can correspond to reduced respect for such systems
	- ↑↓ Extensive cross-functional collaboration?
	    - ↑ Can correspond with flexible systems
	    - ↓ Can indicate complex interdependencies

Government risk	To what extent is there:
	- ↑ Risk of antitrust or regulatory backlash if power is too concentrated?
	- ↑ Pressure from politicians or autocrats to act in a way clearly against the wishes of the public?
	- ↑ Limited oversight?
	- ↑↓ High scrutiny environment?
	    - ↑ Proactive democratization can accelerate preferential regulatory outcomes
	    - ↓ Lack of regulator experience with DA systems may limit influence
	- ↑↓ Complex regulatory landscape?
	- ↓ Clear existing law?

Collective intelligence	To what extent does the decision: - ↑ Benefit from a broad diversity of perspectives? - ↑ Benefit from locally distributed knowledge? - ↑ Involve high levels of uncertainty? - ↓ Involve complex interdependencies?

Figure 6: Full version of the Levels Decision Tool (First Half)

Viability

To what extent:

- ↑↓ Are there political obstacles to delegating power?
  - ↑ Democratic legitimacy could resolve barriers
  - ↓ Political obstacles are despite democratic norms
- ↓ Are there organizational structure, legal, technical, or physical obstacles to further delegating or devolving power?

Resources

To what extent:

- ↑ Is the level of resourcing made available for the democratic system commensurate with the importance of its decisions?
- ↑ Can resources be made available for recurring expenses (only impacts L4, L5)?

Speed

To what extent does the decision involve:

- ↑↓ Time-critical responses?
  - ↑ Can established processes operate within time constraints
  - ↑ L4 processes can respond to regular time-critical scenarios in specified ways
  - ↓ Rules out certain complex systems
- ↓ Emergency responses?

Adaptability

To what extent do you anticipate:

- ↓ Rapid changes to internal or external conditions that might impact the decision, relevance, or remit of a process?
  - ↑ This may be counteracted by more repeating or adaptive L4 or L5 systems

Novelty

To what extent:

- ↑↓ Is devolving such decisions novel?
  - ↑ Democratic first-mover might increase respect and newsworthiness
  - ↓ Exposed to unknown risks

Figure 7: Full version of the Levels Decision Tool (Second Half)