# Toward Democracy Levels for AI

**Aviv Ovadya**[1*] **Luke Thorburn**[1,2] **Kyle Redman**[1,3] **Flynn Devine**[4]
**Smitha Milli**[5] **Manon Revel**[5] **Andrew Konya**[6] **Atoosa Kasirzadeh**[7]

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

aviv@aidemocracyfoundation.org

## Abstract

There is increasing concern about the unilateral power of the organizations involved in the development, alignment, and governance of AI. Recent pilots — 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. However, there is no standard framework for evaluating such processes. In this paper, building on insights from the theory and practice of deliberative democracy, we provide a "Democracy Levels" framework for evaluating the degree to which decisions in a given domain are made democratically. The framework can be used (i) to define milestones in a roadmap for the democratic AI, pluralistic AI, and public AI ecosystems, (ii) to guide organizations that need to increase the legitimacy of their decisions on difficult AI governance questions, and (iii) as a rubric by those aiming to evaluate AI organizations and keep them accountable.

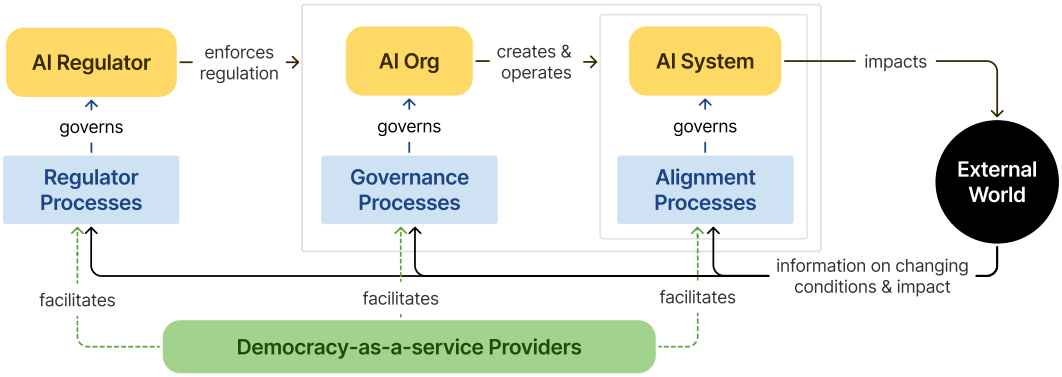



Figure 1: A system diagram of how democratic processes could integrate with the AI ecosystem.

## 1 Introduction

How should we navigate disagreements about the roles and responsibilities of AI systems, AI organizations, and AI regulators in a pluralistic and multipolar world? Similar questions have emerged with previous technological advances [46, 16, 36], 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

Pluralistic Alignment Workshop at NeurIPS 2024.

unprecedented challenges [1]. Thankfully, recent innovations in collective decision-making point towards a new generation of processes, infrastructure, and institutions to navigate these challenges [33, 12, 40].

We focus here on approaches to collective decision-making that are "characterized by a kind of equality among the participants at an essential stage of the decision-making process" [11], which we refer to as **democratic processes**. Beyond the potential normative and pluralistic benefits of such processes, the legitimacy and buy-in engendered by such processes are increasingly valuable to those involved in the development and governance of AI across corporations, governments, and multilateral NGOs [40].

As a result, most leading AI organizations have begun experimenting with such processes for policy or alignment decisions — including Anthropic's *Collective Constitutional AI* [2], OpenAI's *Democratic inputs to AI* grant program [18], Meta's *Community Forums* [8], and Google DeepMind's STELA project [6] — and there is increasing pressure to use such processes for the development of international regulation [13, 33]. While these early steps are clearly imperfect, they are developing internal organizational capacity to understand the potential of this direction, and may act as a stepping stone toward something truly impactful. However, to fulfill the potential of these processes and ensure they are credible, we need a shared language to describe and evaluate progress.

**Contribution.** In this paper, we provide a "Democracy Levels Framework" for understanding the maturity of collective decision-making processes relating to AI, building on innovations from modern deliberative democracy [32] and pluralistic technology [44]. This framework defines a set of levels and dimensions which can:

1. Be used as **milestones in a concrete roadmap** (or "tech tree") for the democratic AI [12], pluralistic AI [39], and public AI [43, 35] ecosystems — a rapidly evolving set of organizations, institutions, and initiatives focused on ensuring that we have the necessary "democratic infrastructure" for navigating the transition to a world with highly-capable AI systems.
2. Help **guide organizations** and institutions that need to increase the legitimacy of their decision-making on difficult AI governance questions.
3. Be used as a **rubric** by those aiming to keep those AI organizations accountable.

We see this framework as being applicable to each of the yellow components in Figure 1: AI systems, AI organizations, and AI regulators (and the decision-making processes that feed into these). The ultimate intent is to provide 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.

**Background.** The framework we describe is agnostic to the kind of democratic process employed for decision-making. However, the examples we provide here generally rely on modern *deliberative* democratic processes [20]. This is due to their ability to work with jurisdictions of arbitrary size, infrastructure, and political structure (including globally), and effectively incorporates the knowledge of diverse participants and subject matter experts [24, 33, 32] (e.g., [30, 19, 17]). A democratic process of this form has as its input a **remit** and **constituent population**, and as its output a **decision**. The remit is a prompt which scopes the decision that needs to be made (and may specify the structure and properties required for the output), and at an essential stage in the process, decisions are made by a representative subset of the constituent population (generally selected by sortition [32]). The process is often conducted by a third-party **democracy-as-a-service provider** [33], analogous to a polling organization, who have expertise in conducting deliberative processes (e.g., [23, 31, 28, 26]).

To develop this framework, we have drawn inspiration from existing frameworks for evaluating democratic-ness (e.g., [3, 25, 22, 38]), as well as frameworks for evaluating degrees of responsible behavior and autonomy in AI systems [7, 37]. Our work relates to explorations and assessments of democratic [12], participatory [15, 14, 41], pluralistic [39], and public AI [35, 43]. We also provide a brief overview of analogous innovation in traditional governments in Appendix A.

## 2 Democracy Levels Framework

**Levels.** We define each level of democratic decision-making according to which of five roles — (i) informing decisions, (ii) specifying options, (iii) making decisions, (iv) initiating decision-making processes, and (v) metagovernance — are performed by democratic processes, rather than a **unilateral authority** (such as a company or some kinds of regulatory bodies). Figure 2 provides definitions of

each level, from Level 0 to Level 5, in these terms, along with concrete examples of what this could look like in practice for a plausible decision domain: a developing a set of rules governing persuasion by an AI system. Such rules might be directly used in model training (for aligning an AI system) [29] or as policies (for an AI organization or regulator).

| | 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** | informing decisions | 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** | informing decisions, specifying options | 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** | informing decisions, specifying options, making decisions | 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** | informing decisions, specifying options, making decisions, initiating processes | The unilateral authority pre-commits to **triggering** 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** | informing decisions, specifying options, making decisions, initiating processes, metagovernance | 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 (potentially 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 such as multi-body sortition).* |

informing decisions · specifying options · making decisions · initiating processes · metagovernance

Figure 2: Definitions of the Democracy Levels.

Note that the processes used for different domains of decision-making might be at different levels. For example, decisions about whether to release a new model might be at Level 2, while decisions about the model spec used for fine-tuning are at Level 4. The remit may also provide scope limitations: for example, specifying that any rules developed via the process would only be binding for two years, or until a given condition is met (such as a model passing a particular benchmark).

**Dimensions.** The *levels* define which roles are performed by democratic processes (versus the unilateral authority). However, in order to meaningfully, safely, and effectively implement the higher democracy levels, there are significant prerequisites that must be met. Those prerequisites can defined in terms of three primary *dimensions*—deliberation, delegation, and trust—each with several sub-dimensions (summarized in Table 1).

First, to level up, the democracy-as-a-service provider must improve the **deliberation** it provides, that is, improve the quality of the democratic processes it facilitates such that it is helpful and safe to shift power to them. More concretely, it must make its processes increasingly *representative* of the relevant population, *informed* in their deliberation, *substantive* in their outputs, *robust* to adversarial behavior and less-than-ideal conditions (e.g., low turnout), and *legible* (transparent and understandable) to non-participants. Second, to level up the unilateral authority must increase its capacity to **delegate** to the democratic process. This includes its capacity to organizationally and publicly *commit* to the outcomes; to *integrate* the capacity to commission and act on such processes into its routine operations; and to technically and/or legally *bind* itself to the resulting decisions. Finally, to level up there must be external conditions which support the success of the process, which we collectively

Table 1: Overview of the dimensions, actors, and sub-dimensions.

| Dimension | Relevant Actor(s) | Sub-dimensions |
|---|---|---|
| Deliberation | Democracy-as-a-service provider | Representation, Substantiveness, Robustness, Informedness, Legibility |
| Delegation | Unilateral authority | Integration, Commitment, Bindingness |
| Trust | Everyone else | Awareness, Buy-in, Participation, Accountability |

refer to as **trust**. Specifically, the relevant public and stakeholders must be sufficiently *aware* of this kind of process, buy into its *legitimacy*, and be willing to *participate*; and there must be sufficiently capable watchdogs for *accountability*.

**Design decisions.** There are a few notable decisions embedded in this framework. We intentionally divorce deliberation from delegation, as third-party democracy-as-a-service providers can be commissioned by unilateral authorities such as AI labs and regulators [10, 34, 8, 42, 2]. This separation of concerns can help prevent fraud and provides an opportunity for process improvements by one organization to be passed on to other organizations. Operationally, at Level 1 such commissioned deliberations are roughly analogous to commissioning representative surveys or community engagement processes, and above that level the processes being commissioned are more sophisticated and more directly integrated into organizational decision-making. These dimensions also have significant dependencies, for example, processes that fail to demonstrate a sufficient quality of deliberation given the level of power entrusted to them (delegation) are likely to lead to backlash (low trust; e.g., if a process is subverted, or a democratic decision sounds good but ends up being counterproductive).

**Examples.** Within this framework, Anthropic's Collective Constitutional AI [2] was arguably an example of a transition from Level 0 to Level 1 (with a few sub-dimensions closer to Level 2). Meta's Oversight Board content decisions are Level 4 across delegation dimensions within a very small remit, but their broader policy recommendations are Level 1 across the same dimensions—and both are below Level 1 for representation [27]. Meta's Community Forum on AI conversely was above Level 1 for representation [10]. For each dimension it is possible to specify in detail what it takes to level-up, and use that to guide improvement and investment. We provide initial examples in Appendix B.

## 3 Limitations & Discussion

Our framework for evaluating democratic decision-making focuses only on a general notion of democracy and makes a number of simplifying assumptions in order to provide an accessible map. The specific context, such as the kind of decision and the extent of polarization around it, can impact the quality of deliberation required. Moreover, this framework does not provide guidance on when a higher democracy level is warranted, given the costs and risks; when a decision should delegated directly to users to decide for themselves (i.e., if there are minimal externalities), or what the constituent population for a decision should be (e.g., local, national, global).

This framework can also be used to clarify when organizations are claiming to be acting more democratically than they actually are—and that can then help provide a basis for ensuring that they live up to their professed standards. This differential between ambitious democratic aspirations and reality has been a major force for democracy across history. Maturity also doesn't come overnight—organizations, democracy-as-a-service 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 their current democracy level or status quo alternatives), both of which can be articulated through a level system.

By providing a concrete articulation that may be contested and built upon, we hope that this framework 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.

## Acknowledgements

The authors would like to thank Danielle Allen, Nora Ammann, Joe Edelman, David Evan Harris, Sarah Hubbard, Jonas Kgomo, Evan Shapiro, Allison Stanger, Jessica Yu, Kinney Zalesne, and others, for feedback on earlier versions of this paper.

Luke Thorburn was supported in part by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (safeandtrustedai.org), King's College London.

## References

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

[2] Anthropic. Collective Constitutional AI: Aligning a Language Model with Public Input, 2023. URL https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input. Accessed: 2024-09-03.

[3] Sherry R. Arnstein. A ladder of citizen participation. *Journal of the American Institute of Planners*, 35(4):216–224, 1969. doi: 10.1080/01944366908977225. URL https://doi.org/10.1080/01944366908977225.

[4] Assemblée citoyenne pour le climat. Bruxelles en 2050, vous la voyez comment ?, 2024. URL https://www.assembleeclimat.brussels/.

[5] Barwon Water. Water for our Future Community Panel update, 2021. URL https://www.barwonwater.vic.gov.au/about-us/news-and-events/news/Water-for-our-Future-Community-Panel-update.

[6] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. STELA: A community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14(1):6616, 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-56648-4. URL https://www.nature.com/articles/s41598-024-56648-4. Accessed: 2024-09-04.

[7] Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The Foundation Model Transparency Index v1.1: May 2024, 2024. URL http://arxiv.org/abs/2407.12929. Accessed: 2024-09-04.

[8] Jennifer Broxmeyer. Leading the Way in Governance Innovation With Community Forums on AI, 2024. URL https://about.fb.com/news/2024/04/leading-the-way-in-governance-innovation-with-community-forums-on-ai/. Accessed: 2024-09-03.

[9] Bürgerdialog in Ostbelgien. What is the Citizens' Council?, 2024. URL https://www.buergerdialog.be/en/citizen-council/what-is-the-citizens-council.

[10] Samuel Chang, Estelle Ciesla, Michael Finch, James S. Fishkin, Lodewijk L. Gelauff, Ashish Goel, Ricky Hernandez Marquez, Shoaib Mohammed, and Alice Siu. Meta Community Forum Results Analysis. Technical report, Stanford Deliberative Democracy Lab, 2024. URL https://cddrl.fsi.stanford.edu/publication/meta-community-forum-results-analysis.

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

[12] Collective Intelligence Project. A Roadmap to Democratic AI. Technical report, The Collective Intelligence Project, 2024. URL https://cip.org/s/CIP_-A-Roadmap-to-Democratic-AI.pdf.

[13] Connected by data. Options for a Global Citizens Assembly on AI, 2024. URL https://connectedbydata.org/projects/2024-gca-ai. Accessed: 2024-09-04.

[14] Ned Cooper and Alex Zafiroglu. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2024-05-11. doi: 10.1145/3613904.3642775. URL http://arxiv.org/abs/2403.06431. Accessed: 2024-09-04.

[15] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice, 2023. URL http://arxiv.org/abs/2310.00907. Accessed: 2024-09-04.

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

[17] Deutscher Bundestag. Final Recommendations of the Citizens' Assembly on Nutrition, 2024. URL https://www.bundestag.de/en/parliament/recommendations-nutrition-995916.

[18] Tyna Eloundou and Teddy Lee. Democratic inputs to AI grant program: Lessons learned and implementation plans, 2024. URL https://openai.com/index/democratic-inputs-to-ai-grant-program-update/. Accessed: 2024-09-03.

[19] European Commission. European Citizens' Panel on Tackling Hatred in Society - Final Recommendations, 2024. URL https://citizens.ec.europa.eu/european-citizens-panels/tackling-hatred-society-panel_en.

[20] James S. Fishkin. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford University Press, 08 2011. ISBN 9780199604432. doi: 10.1093/acprof:osobl/9780199604432.001.0001. URL https://doi.org/10.1093/acprof:osobl/9780199604432.001.0001.

[21] G1000. Citizens' panel on the challenges of aging, 2017. URL https://www.g1000.org/en/cases/citizens-panel-challenges-aging.

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

[23] Involve. Involve, 2024. URL https://www.involve.org.uk/. Democratic Infrastructure Provider.

[24] Helene E. Landemore. Why the Many Are Smarter than the Few and Why It Matters. *Journal of Deliberative Democracy*, 8(1), 2012. ISSN 2634-0488. doi: 10.16997/jdd.129. URL https://delibdemjournal.org/article/id/401/. Accessed: 2024-09-04.

[25] Staffan I. Lindberg, Michael Coppedge, John Gerring, and Jan Teorell. V-Dem: A New Way to Measure Democracy. *Journal of Democracy*, 25(3):159–169, 2014. ISSN 1086-3214. doi: 10.1353/jod.2014.0040. URL https://muse.jhu.edu/article/549506. Accessed: 2024-09-04.

[26] MASS LBP. MASS LBP, 2024. URL https://www.masslbp.com/. Democratic Infrastructure Provider.

[27] Meta Oversight Board. Oversight Board Charter. https://www.oversightboard.com/wp-content/uploads/2024/03/OB_Charter_March_2024.pdf, 3 2024.

[28] Mosaic. MosaicLab, 2024. URL https://www.mosaiclab.com.au/. Democratic Infrastructure Provider.

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

[30] Nakyung Lee, Susan and John M Stormonth-Darling and Granaz Baloch and Yago Bermejo Abati and Flynn Devine and Sow Ebion, Eva and Islam Elbeiti and Brett Henning and Marius Kamugisha and Jamie Kelsey-Fry and Claire Mellier and Jon Stever and van der Stoep, Remco and Sarah Whitley and Rich Wilson and Fatima Zamba. Report of the 2021 Global Assembly on the Climate and Ecological Crisis. Technical report, Global Assembly, 2022. URL https://globalassembly.org/resources/downloads/GlobalAssembly2021-FullReport.pdf.

[31] Nexus. Nexus Institute, 2024. URL https://nexusinstitut.de/en/. Democratic Infrastructure Provider.

[32] OECD. *Innovative Citizen Participation and New Democratic Institutions*. OECD iLibrary, 2020. doi: https://doi.org/https://doi.org/10.1787/339306da-en. URL https://www.oecd-ilibrary.org/content/publication/339306da-en.

[33] Aviv Ovadya. Reimagining Democracy for AI. *Journal of Democracy*, 34(4):162–170, 2023. ISSN 1086-3214. doi: 10.1353/jod.2023.a907697. URL https://muse.jhu.edu/article/907697. Accessed: 2024-09-04.

[34] Aviv Ovadya. Meta Ran a Giant Experiment in Governance. Now It's Turning to AI. *Wired*, 2023. URL https://www.wired.com/story/meta-ran-a-giant-experiment-in-governance-now-its-turning-to-ai/.

[35] Public AI Network. Public AI: Infrastructure for the common good. Technical report, The Public AI Network, 2024. URL https://publicai.network/whitepaper.

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

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

[38] Svend-Erik Skaaning and Alexand Hudson. *The Global State of Democracy Indices Methodology: Conceptualization and Measurement Framework, Version 7 (2023)*. International Institute for Democracy and Electoral Assistance (International IDEA), 7 edition, 2023. ISBN 978-91-7671-644-1. doi: 10.31752/idea.2023.38. URL https://idea.int/sites/default/files/GSOD/global-state-of-democracy-indices-methodology-v7-2023.pdf. Accessed: 2024-09-04.

[39] 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. A Roadmap to Pluralistic Alignment, 2024. URL http://arxiv.org/abs/2402.05070. Accessed: 2024-09-04.

[40] Jack Stilgoe. Ai has a democracy problem. citizens' assemblies can help. *Science*, 385(6711): eadr6713, 2024. doi: 10.1126/science.adr6713. URL https://www.science.org/doi/abs/10.1126/science.adr6713.

[41] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1609–1621, Rio de Janeiro, Brazil, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658992. URL https://doi.org/10.1145/3630106.3658992.

[42] The Forum for Ethical AI. Democratising Decisions About Technology: A Toolkit . Technical report, The RSA, 2019. URL https://www.thersa.org/globalassets/reports/2019/democratising-decisions-tech-report.pdf.

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

[44] E. Glen Weyl, Audrey Tang, and "Plurality Community". *Plurality: The Future of Collaborative Technology and Democracy*. Independently Published, 2024. ISBN 979-8321247181. URL https://www.plurality.net/chapters/.

[45] Yarra Valley Water. Yarra valley water price submission 2017, 2017. URL https://www.yvw.com.au/news-room/news-archive/our-submission-pricing.

[46] Malte Ziewitz and Ian Brown. *A prehistory of internet governance*, chapter 1, pages 3–26. Edward Elgar Publishing, Cheltenham, UK, 2013. ISBN 9781849805025. doi: 10.4337/9781849805025.00008. URL https://www.elgaronline.com/view/edcoll/9781849805025/9781849805025.00008.xml.

# Appendices

## A Context from non-AI governance

In developing this framework, we also draw on experience with increases in democratic maturity in government. For example, the emergence of modern representative democracy in the United Kingdom evolved over time from ad-hoc parliaments called on demand by the monarchy for decisions that needed significant buy-in, to institutionalized bodies within a larger system of checks and balances, where the monarch has negligible power.

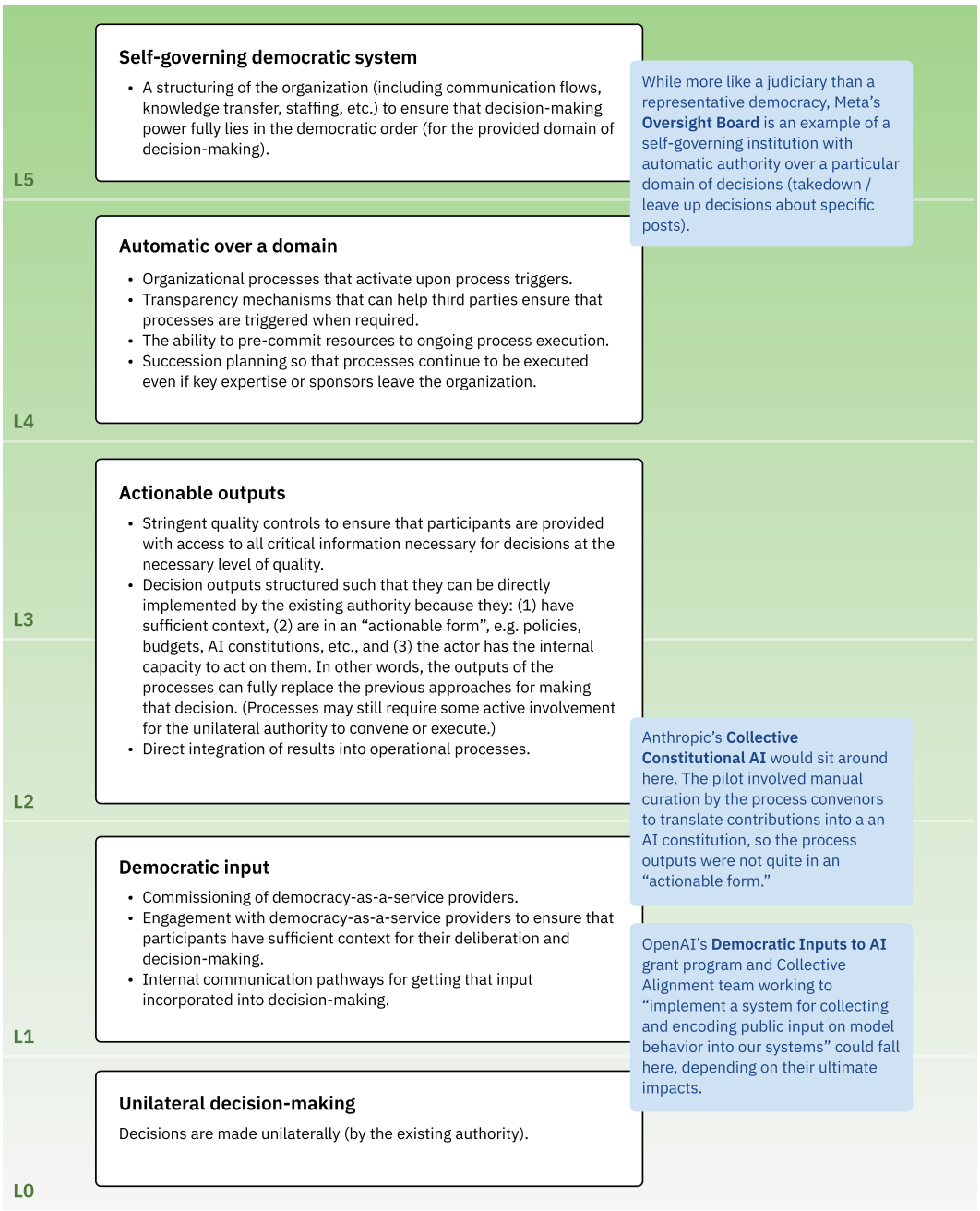
More recently, governments and organizations have increasingly integrated deliberative democracy processes when executive bodies needed to balance challenging tradeoffs or increase legitimacy. Some state-owned water businesses in Australia have, over the past seven years, begun to make long-term business plan decisions using citizens' juries (e.g., Yarra Valley Water [45], Barwon Water [5]). In Belgium, regional governments have built on ad-hoc citizens' panels (e.g., Citizens' Panel on the Challenges of Aging [21]) and have now established permanent citizens' bodies (e.g., Ostbelgien Citizens' Council [9], Brussels Climate Assembly [4]). For each of these cases, we can map out the transitions as democratic processes became increasingly core to their decision-making.

## B Detailed Scales for the Dimensions

For each of the dimensions of the framework introduced in Section 2, there is ongoing work towards providing a detailed set of incremental steps that can guide improvements at the margin and be used for granular evaluation. Below, we a provide rough sense of what this might look like for the delegation dimensions, and give examples of where existing processes might fall on these scales. Similar scales for the other dimensions will be included in a future version of the framework.

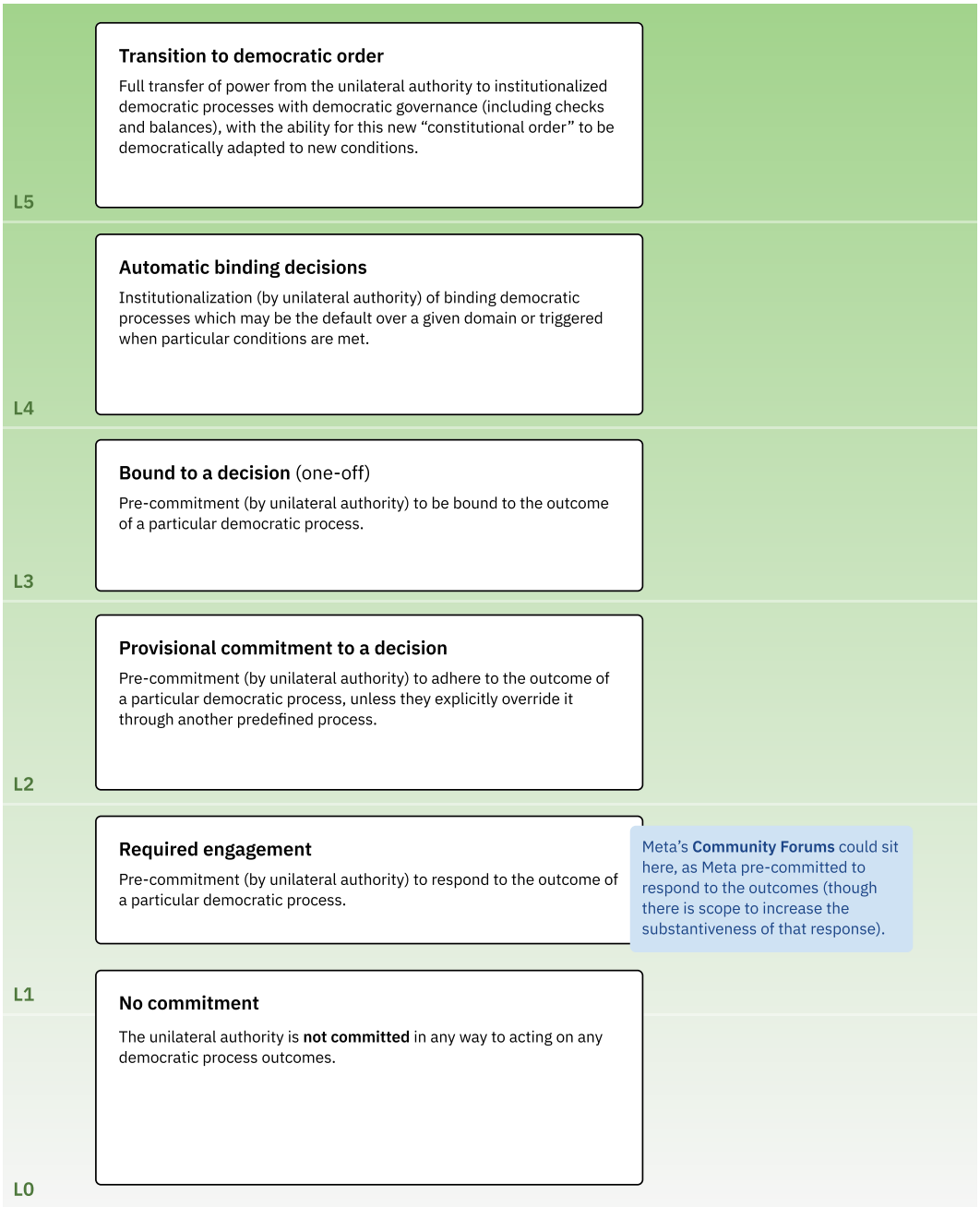
### B.1 Integration

This dimension captures the degree to which an actor's decision-making processes can be effectively integrated with democratic processes.



## B.2 Commitment

This dimension captures the degree to which an actor commits itself to acting in a way that is consistent with the outputs of a democratic process.



### B.3 Ability to bind

This dimension captures the degree to which an actor can meaningfully bind itself to the decision of a democratic process, should it choose to do so.

L5

L4

L3

**Able to bind**

The unilateral authority can meaningfully bind itself to the decisions of a democratic process. All primary blockers to pre-commitment have been overcome.

(Caveat: there will always be fundamental limitations, such as those relating to physical and mathematical limits or some kinds of superseding jurisdictional law.)

**Anthropic** is a Delaware Public Benefit Corporation (PBC), and Delaware corporate law "expressly permits the directors of a PBC to balance the financial interests of the stockholders with the public benefit purpose specified in the corporation's certificate of incorporation," perhaps by binding itself to democratic processes (though there may be other barriers).

L2

L1

L0

**Not able to bind**

The unilateral authority cannot meaningfully bind itself to the decisions of a democratic process, due to any of:

- countervailing incentives,
- legal structure (e.g., fiduciary duty for some kinds of organizations such as a Delaware C Corp),
- lack of capacity to enforce or operationalize decisions, or
- technical challenges (e.g., for binding/aligning AI systems to goals or constraints).