

Process Description

AI Democracy

Landscape Map

The Power for Democracies' AI-Democracy Landscape Map sets out to provide the first comprehensive view of threats, mitigations, and opportunities at the intersection of emerging technology and democracy. Here, the research lead explains how the map was developed, and outlines the analyses the database enables, as well as how to interpret the data and frameworks.

Author: Karoline Helbig, Senior Researcher and Project Lead

Support: Aurora Mane, Research Fellow

August 2026

© 2026 Power for Democracies gGmbH
All rights reserved.

P4D_2026_PRJ02_RP002

Power for Democracies strengthens and protects democracy by identifying and researching and recommending highly effective civil society initiatives for donors to support. It is non-profit and non-partisan. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Power for Democracies.

This publication received no specific grant from any funding agency or donor.

Registered office:
c/o UES – Gemeinnützige GmbH für Effektives Spenden
Müllerstrasse 138D
13353 Berlin, Germany

hello@powerfordemocracies.org
www.powerfordemocracies.org
+49 30 12144574

Executive Director: Markus N. Beeko

Content

<u>Introduction</u>	4
<u>Step 1: Source selection and literature review</u>	5
<u>Step 2: Extraction and standardisation</u>	6
<u>Step 3: Democracy aspects coding</u>	7
<u>Step 4: Harm mechanism coding</u>	13
<u>Step 5: Pro-democracy activities coding</u>	16
<u>What the coded database makes possible</u>	17
<u>Interpretive notes and limitations</u>	19
<u>Appendix</u>	24
<u>Democracy Aspects Codebook</u>	24
<u>Code Overview</u>	25
<u>Code Descriptions (Detailed)</u>	26
<u>Harms Mechanisms</u>	35
<u>Pro-Democracy Mechanisms</u>	49
<u>References</u>	55

Introduction

When Power for Democracies set out to explore the impact of AI on democracy as a potential field of research from which to produce donor recommendations, we found a disparate body of work on the subject. While several pieces of research and explorations existed, they were scattered across a wide spectrum of topics, and differed in scope, granularity, and perspective. The first step therefore was to create a map of the existing landscape to draw insights into what is known about threats, mitigations, and opportunities.

We built a structured database that maps AI threats to democracy, the strategies that have been identified to address them, and the opportunities AI creates for democratic resilience. As of June 2026, it contained 161 documented threats, 117 pro-democracy strategies, and 43 independent pro-democracy opportunities. The database is designed to grow as new sources are identified and our analytical frameworks develop.

Our process has so far consisted of five main phases: (1) selecting and reviewing source material; (2) extracting and standardising entries into a unified database; (3) coding each entry against a democracy aspects framework; (4) coding each entry against a tiered harm-mechanism framework; and (5) coding each entry against a pro-democracy activities framework. Our process was iterative: we tested approaches, revised them, and, in some, cases started over.

This collation of the literature represents the beginning of a comprehensive map of the field for actors seeking an introduction into the intersection of AI and democracy. It may be useful, for example, for civil society designing new programmes, grant-makers exploring new funding streams, and researchers seeking an overview of the field.

This process description sets out each of these phases in turn, before closing with two sections on what the coded database makes possible and what to bear in mind when interpreting figures drawn from it.

Step 1: Source selection and literature review

We began by identifying existing frameworks, reports, and analyses that systematically mapped AI-related effects on democracy. Our goal was to synthesise the findings from existing research and to surface patterns that remain invisible when sources are read in isolation, rather than to conduct primary empirical research from scratch.

Through desk research and expert interviews, we sought out sources that had already catalogued or systematically identified AI threats to democratic processes, potential mitigation strategies, or pro-democracy opportunities.

We reviewed a larger initial body of literature and selected ten sources¹ for the pilot phase of the database. For this initial selection, we prioritised:

- **breadth across democratic dimensions and threat types** – sources that included a wide range of ways AI can affect democracy and not only the most visible ones.
- **disciplinary variety** – the selection spans formats and methodologies, including policy surveys, electoral integrity analyses, AI safety frameworks, and academic political science research.
- **variation in granularity level** – the selection includes sources covering high-level threat categories, as well as those which involve technical detail about specific attack vectors.
- **relevance to democratic effects specifically** – excluded were sources focused on AI risk in the abstract where the connection to democratic processes was not made explicit.

¹Bengio et al., 2026; European Parliament, 2025; George et al., 2026; Guzman Piedrahita et al., 2026; Jungherr, 2023; Kela, 2025; Ovadya et al., 2025; Panditharatne et al., 2025; Tabassi, 2023; Tsai et al., 2024). Guzman Piedrahita et al., 2026 was a pre-print at the time, and we coded the version dated 23 March 2026, and it has received updates since. European Parliament, 2020 is a continuously updated web page with its last update published in April 2025.

This selection is preliminary and limited.² Sources originate from Europe and the United States,³ although some sources are authored by international groups. Moreover, several strong sources have been published or identified since our initial selection that would enrich the database, including some that cover dimensions that are currently thin or missing in our source set. We hope to revisit and expand the source selection in subsequent phases.⁴

Step 2: Extraction and standardisation

We extracted entries from all ten sources and standardised them into a unified database. Each entry represents either a discrete threat for which no mitigation strategy has been explicitly identified in the source text; a threat-and-mitigation-strategy combination as presented in the source text; or an independent pro-democracy opportunity identified in the source text without being matched to a threat. For each entry, we:

- selected a verbatim quotation from the text for each threat, mitigation strategy, and opportunity.
- wrote a paraphrase to give minimal additional context where needed.
- noted the source.

We used AI to assist with the extraction of material, and two human coders verified the results. This process required some judgement calls to be made. Due to the considerable variation of the sources in form, scope, granularity etc. we aimed to maintain consistency by extracting entries at a level of specificity sufficient to support meaningful coding, but not so

² One source in particular sits at some distance from the project's core focus. It is primarily a cybersecurity document, with limited direct engagement with democratic processes. We retained it because it offers substantially more technical depth on the mechanics of AI-enabled fraud and deception than our other sources. This gives us some analytically useful detail for understanding how specific threat mechanisms work in practice. We are aware however that this creates some unevenness in the set and we hope to address this gap with the future addition of other literature.

³ We looked for sources that were the most comprehensive and widely cited at the time, in addition to meeting the selection criteria above. The Western focus of the resulting set became apparent during this process.

⁴ We hope to add central and widely cited sources (e.g. Amodei, 2026; Costa, 2026; Kerley, 2023; Kerley et al., 2024; Kulveit et al., 2025; Summerfield et al., 2025).

granular as to fragment phenomena that the sources treated as unified or connected.

During the subsequent coding passes, particularly for the harm mechanisms, some entries were further split to accommodate different mechanisms. We also needed to classify text passages as threat, mitigation strategy, or independent opportunity. This classification was mostly discernible from the text, but in rare cases closer engagement with the data revealed some miscoding that had to be adjusted during coding passes for democracy aspects or mechanisms.

The resulting map is an interactive table, searchable by type of impact (threat, threat-mitigation pairing, independent opportunity) and by source. For each entry, the verbatims can be extracted with their respective sources.

Step 3: Democracy aspects coding

We coded each database entry against an adapted democracy aspects framework – a category system that systematises the different processes, spaces, and institutions integral to democratic systems. This layer of coding identifies which dimension or dimensions of democracy a given threat, strategy, or opportunity affects. It enables users to ask, for instance, which democratic domains are most exposed to AI-related threats, or which threats are rarely or not matched with mitigation strategies in the original sources.

We adapted our framework from the International IDEA Democracy Assessment Framework (Beetham et al., 2008). IDEA's framework, a practitioner guide built around assessment questions for national democratic audits, provided a strong foundation and structure: four pillars and a set of subcategory names. While the IDEA framework is built to assess the democratic quality of countries, ours is designed to identify the aspects of democracy that AI use affects, whether positively or negatively. We adapted the framework for this purpose by adding category definitions, conceptual

explanations, and coding rules to enable systematic analysis of our diverse document corpus.

For the Power for Democracies' version of the democracy aspects framework, we developed the definitions, conceptual rationale for each category, decision rules for ambiguous cases, and worked examples. We largely retained IDEA's pillar structure and category names, with some adjustments, including adding subcategories, reframing existing categories, and slightly restructuring the framework.

Our changes	Why
All categories defined and refined	We extrapolated definitions from IDEA's questionnaire items, refined them towards AI-focused applicability, renamed them for better accessibility, and developed an independent codebook.
2.5 Judiciary added	We added a judiciary-focused category based on the original categories 1.2 and 2.3.
2.8 Illegitimate influence renamed and sharpened	We reframed IDEA's 2.6 'Integrity in Public Life', which focuses primarily on corruption and bribery, to capture AI-enabled back-channel influence and manufactured consensus.
3.1 Media narrowed	We narrowed this category to align more closely with the deliberative-systems approach, and to improve its fit with the AI context.
3.2 Added will-formation	Additional to IDEA's 'Political Participation' involving civil society and voter turnout, we extended the category to include opinion formation and the broader digital information environment, again drawing on the deliberative-systems approach as a better fit for the AI context.
Several categories renamed and one relocated	We renamed IDEA's 'Democracy Beyond the State' as 'transnational dynamics', and 'Effective and Responsive Government' as 'effective and accountable government', to make them more immediately comprehensible. We moved IDEA's 'Decentralization' from pillar 3 into pillar 2 as 'national-subnational power relations', to situate it along with the other questions about the distribution of state power.

The adjustments are based on theoretical considerations and our understanding of how AI affects democracy. The IDEA framework was designed primarily within a liberal-democratic and institutionalist tradition. AI's effects on democracy, as indicated in our map entries, called for adjustments in two directions.⁵

First, we gave greater weight to deliberative and participatory democratic theory (Parkinson and Mansbridge, 2012). We reframed both civil society-related categories to better capture the conditions under which citizens form political views, participate in deliberation, and exercise collective agency.⁶ Since a substantial share of AI's impact on democracy concerns these participatory and deliberative processes, we considered these adjustments to be useful.

Second, we extended the framework to capture AI's effects on the structural distribution of power. This had only been partially represented in the IDEA framework as perceived corruption. The concentration of power around AI companies due to the gatekeeper role they may exercise over both the information environment and as infrastructure provider for a wide range of government functions – such as knowledge processing, social service provision, election management, management of physical infrastructures – is qualitatively new and thus not captured in the original framework. We therefore reframed the original corruption category (now 'illegitimate influence over policymaking') to account for these dynamics.

Our adjusted framework structures democracy aspects into four pillars and 16 subcategories:

⁵ Moreover, we highlighted the importance of checks and balances by adding a code for the judiciary, thereby giving each pillar its own subcategory. This was, however, a conceptual decision rather than one driven by the impact of AI on political systems.

⁶ This also meant that items referring to social media are now coded under 'opinion formation and political participation' rather than the 'media' code.

1 Citizenship, law, and rights		
1.1	Nationhood and citizenship	Effects on inclusion, equal recognition, or protection of minorities and marginalised groups.
1.2	Rule of law and access to justice	Effects on legal processes, due process, and citizens' ability to obtain fair treatment and redress.
1.3	Civil and political rights	Effects on freedoms of expression, assembly, or association, including through surveillance or censorship.
1.4	Economic and social rights	Effects on access to work, material conditions, or distribution of economic power.
2 Representative and accountable government		
2.1	Free and fair elections	Effects on electoral integrity, voter access, campaign fairness, or trust in election outcomes.
2.2	Democratic role of political parties	Effects on political competition, party financing, or the balance between parties.
2.3	Effective and accountable government	Effects on transparency, accountability, or fairness of government services and policymaking.
2.4	Democratic effectiveness of legislature	Effects on legislative scrutiny, oversight capacity, or executive-legislature power balance.
2.5	Democratic role of the judiciary	Effects on judicial independence, constitutional oversight of government, and citizens' ability to challenge the exercise of public power through legal means (strategic litigation).
2.6	National–subnational power relations	Effects on the distribution of power between central and local governments.
2.7	Civilian control of military and police	Effects on oversight and accountability of security and enforcement agencies.
2.8	Illegitimate influence over policymaking	Effects on corruption, regulatory capture, or undue private influence over public policy and the policy process.
3 Civil society and popular participation		
3.1	Independent and fact-based journalistic media	Effects on the independence, viability, and reach of regulated, professionally produced journalism.

3.2	Opinion formation and political participation	Effects on how citizens form and exchange political views, participate in public life, or organise through civil society.
4	Transnational dynamics	
4.1	External influences on democracy	Effects on foreign interference or the democratic-autocratic geopolitical balance.
4.2	Democratic impact abroad	Effects of a country's AI policies or exports on democracy in other countries.

We developed a detailed coding protocol to ensure consistency in coding decisions across different coders. A main challenge was that some of the source texts did not directly address democracy, while many used wording that was different from that of our framework.

We coded a democracy aspect where the source text names it explicitly, where the verbatim extract describes something conventionally understood to bear on that aspect even if it is not named, and where a downstream effect on it follows clearly from what the source describes. We deliberately stopped short of coding more distant secondary effects: most threats and pro-democracy strategies have a wide variety of downstream consequences, and the system would lose meaning if every entry were coded to every category. As a result, we rarely coded more than four categories for a single entry.

Finally, the source texts were not intended to be disentangled from their wider context,. As a result, not all the quotes – even when considered alongside the paraphrases – captured the full meaning intended in the original text. This could lead to under-coding of some entries.

Before finalising the codebook, we had it reviewed by at least two researchers. We tested all coding categories with two to three human coders and used AI-assisted coding review as an additional consistency check. Where coders disagreed, we worked through the disagreement together and refined the relevant decision rules in the codebook. The full codebook, including definitions, examples, and decision rules for all 16 sub-categories, is available in the [appendix](#).

Step 4: Harm mechanism coding

While the democracy aspects framework establishes which aspect of democracy is affected by a threat, the harm mechanism coding addresses what is actually happening when democracy is threatened. Initially, we wanted to cluster the threats in our datasets to make them more accessible, so we started by inductively developing a threat categorisation system. Rather than deriving categories from an existing theory of technological harm, we worked from what our sources actually describe, grouping recurring mechanisms accordingly.

But in doing so, we realised that a single threat to democracy usually involves several mechanisms, and that these mechanisms are often linked causally: one enables the next, so they can be arranged in chains running from upstream to downstream. Our source texts vary in how much of these chains they capture: some name a single mechanism; others list several; and some describe two or three consecutive steps within a single passage.

These mechanisms are thus highly heterogeneous. Electoral integrity, for instance, is put at risk by fabricated political content, by biased AI use inside election administration, and by cyberattacks on election infrastructure – three problems with little in common in terms of who could address them or how. Conversely, a single mechanism often threatens several democratic values at once, so that countering it would have positive effects across various democracy aspects. Coding for democracy aspects alone cannot surface either pattern. It tells us that elections are frequently discussed, but not how exactly the negative effects are taking hold, and therefore does not bring us closer to identifying possible levers for mitigating the threats.

Distinguishing the mechanisms makes it possible to identify the points along these pathways at which action could be taken. It also shows us which mechanisms are frequently discussed in the analysed literature and how they are connected. We therefore structured the mechanisms in a tiered system that reflects their position in the causal harm pathways. The tiers run from the conditions that shape how AI

systems are built and released, through the properties of the systems themselves and what they are used for, to cumulative downstream effects on individuals and society as a whole.⁷

The framework now comprises eight tiers:

T0	Conditions of AI development	Conditions and choices that shape how AI systems are built, released, and used.
T1	AI capabilities and system properties	Architectural and design features of the AI systems themselves, that have harm potential prior to any specific output or deployment context.
T2	AI output properties	Mechanisms at the level of what individual AI outputs do: the qualities of the artefacts a model produces, realised once an output is generated.
T3	Deployment decisions	Mechanisms that reify with the decision to put AI into use, holding for whoever adopts a system by virtue of the adoption itself.
T4	Basic operations using AI	The elementary operations an AI model can perform, taken independently of the purposes they serve and the effects they produce.
T5	Problematic use of AI	Purposive, real-world application of the elementary operations aimed at an effect on the social world.
T6	Effects on individuals	Mechanisms by which sustained interaction with AI reshapes how individual humans think, decide, and act.
T7	Society-level dynamics	Society-level reshaping of the information environment, public discourse, economic structures, and political-corporate power positions, emerging from many AI uses rather than single events.

⁷ A second finding from the initial inductive categorisation was that democratic harms from AI are not only the product of deliberate misuse. Rather, they frequently arise from individually benign or authorised activity, like citizens using AI to write submissions to a consultation, an agency adopting an AI tool for casework, or a developer choosing an optimisation objective. The tiered system helped capture these harm pathways, even when no actors with nefarious intentions were involved.

Tiers subdivide into subclusters and individual mechanisms, and we coded every entry that comprises a threat verbatim to the most detailed levels. The detailed description of the entire system can be found in the [appendix](#). The framework is not closed: it was derived from the sources in the current selection, and we expect to extend it as further literature is added that describes mechanisms that our present set does not cover.

Two features of the design are worth noting. First, the tiers are causally ordered but not mutually exclusive: most entries carry mechanisms at several tiers, because entries often capture parts of a chain or list parallel mechanisms rather than focusing on a single link. Second, the framework deliberately spans the full range of causal mechanisms from structural conditions to acute misuse, so a regulatory gap and a deepfake campaign can both be located within one system.

Codes were applied only where the verbatim extracts explicitly mentioned the mechanisms, with inferences largely avoided. The coding itself was conducted by two human coders who first completed individual coding passes. We then used AI to flag missed categories or coding discrepancies in the coding of similar threats before the coders coordinated to resolve inter-coder discrepancies. The final coding decision for each entry was made by a human.

The process was highly iterative: throughout the process, we extended and refined the coding system as new mechanisms emerged.⁸ After finalising the harm-mechanism framework, we conducted a final coding pass across the database to ensure that entries were coded against the completed version of the system.

⁸ Due to those iterations, the code numbering is not continuous and the harm mechanism system contains no T2.4. These inconsistencies emerged, where categories were merged, split or dropped as coding proceeded. We retained the original numbers rather than renumbering, so that codes remain comparable across versions of the database. For the same reason, the AI–Democracy Landscape identifies mechanisms and activities by name rather than by code wherever possible.

Step 5: Pro-democracy activities coding

Alongside the harms, we coded each entry containing a pro-democracy verbatim (either a mitigation strategy or an opportunity) against a second system capturing pro-democracy activities. This second system captures activities that strengthen democratic resilience in the face of AI through behavioural adaptation, governance, pro-democracy uses of AI systems, and the evidence base that informs these activities. As with the harm-mechanism framework, we built this system inductively, grouping the activities proposed in our sources rather than deriving categories from an existing theory of democratic resilience.

As with the harm mechanisms, we initially intended to simply cluster the mitigations and opportunities to make them more accessible. Our first attempt to do so led to several findings. First, the available information in our source texts was considerably thinner and less complex for pro-democracy activities than for harms. While some texts link different pro-democracy activities to one another – for example, where a developer-side measure creates the precondition for a downstream activity – these relationships are a minority, are not consistently directional, and are not sufficiently in-depth to allow for a systematic ordering. We therefore clustered the activities by type rather than imposing causal chains which would have required us to supply detail that was not contained in the sources.

Second, the inductive coding of mitigations and opportunities largely produced overlapping coding systems. We therefore collapsed them into one overall pro-democracy activities system.

Third, these activities are undertaken by specific actors and particular activities are overwhelmingly connected with particular actors. For example, responsible development practices are undertaken by developers, government adaptation by state actors, and citizen-led content integrity

efforts by civil society and platforms.⁹ The coding system therefore naturally distinguishes between activities and, at the same time, the actors associated with them.

Fourth, a substantial share of our pro-democracy verbatims addressed more than one category. We therefore decided to apply multiple codes to these entries. The system so far comprises ten [clusters](#)¹⁰

As with the harm mechanisms, several clusters are subdivided into subclusters at which coding was performed. For example, legal frameworks were divided into operational requirements, prohibitions, liability, and market-shaping rules; participation infrastructure into discourse moderation, input synthesis, and participatory formats, etc. A detailed description of the entire system is provided in the [appendix](#). The system is likewise not closed: we expect to extend it as further literature identifies activities that are not captured by our current framework.

Again, codes were applied only when the verbatim extracts explicitly mentioned the activities, with inferences largely avoided. The coding of the pro-democracy activities used the same process as for the coding of the harm mechanisms: it was conducted by two human coders who first completed individual coding passes. We then used AI to flag missed categories or coding discrepancies in the coding of similar entries before the coders coordinated to resolve inter-coder discrepancies. Again, the final coding decision for each entry was made by a human. The coding system was continuously developed and refined, with a final coding pass conducted once the framework had been finalised.

What the coded database makes possible

The three code systems – democracy aspects, harm mechanisms, and pro-democracy activities – are applied to the same entries.

⁹ Harm mechanisms, by contrast, describe processes that can be set in motion by many different actors, and the system is built so that the same mechanism can be coded regardless of who is behind it.

¹⁰ B8 was abandoned and/or merged with another category during the iterative development of this system.

Even when considered individually, each layer already allows for some analysis. The *democracy aspects* coding shows which aspects of democratic systems are discussed most and least frequently in our source texts. The *pro-democracy activities* coding shows which measures to support democracy are mentioned in the sources and, by implication, which actors the literature addresses. The *harm mechanism coding* allows for the most analytical depth. It can show which mechanisms are mentioned most frequently and, because it was developed inductively, provides a sense of the variety of mechanisms mentioned across the sources.

The tiered structure of the harm mechanisms also allows the coding to show how mechanisms are connected in the sources. Because entries commonly carry mechanisms at several tiers, and the tiers are causally ordered, recurring multi-tier combinations can be read as processes the literature describes: from an enabling condition or capability, through the uses it enables, to downstream effects on individuals or society. A tier-based analysis can also show where the sources' focus lies within these causal chains, as well as where they provide less information.

Because the code systems are applied to the same entries, they can also be combined for more in-depth analysis. In particular, combining the harm mechanisms and pro-democracy activities makes it possible to address a wide range of questions. Since both systems are applied to the same entries, each proposed activity can be analysed in relation to the harms alongside which it is discussed, and each harm mechanism in relation to the activities proposed in response to it. Analyses can address questions such as:

- Which pro-democracy activities are mentioned in connection with which harm mechanisms?
- Which pro-democracy activities are mentioned as countering the widest set of harms?
- Which harms can be countered by the largest range of pro-democracy activities, according to our sources?

In combining these coding systems, the absence of connections can be informative as well: some pro-democracy activities are presented independently of any identified threat. They are framed as opportunities to strengthen democracy in the context of AI, rather than as responses to threats posed by AI. Conversely, a large share of entries contains threat coding but no pro-democracy activity coding. These entries identify

harms for which the source texts do not identify a corresponding mitigation strategy to counter them.

A further set of questions connects the democracy aspects framework with the other coding systems. Analyses can address questions such as:

- Through which pathways does the literature link AI to harm for democracy?
- Which democracy aspects are most frequently discussed as being at risk from AI-induced harm?
- Which mechanisms are presented as harming democracy?
- Which pro-democracy activities are frequently linked to which harm mechanisms?

These questions make it possible to distinguish routes to the same democratic harm that have little in common. They also offer a way to identify points at which action could be taken, that is specific harm mechanisms that could be countered within a pathway leading to harm to a particular aspect of democracy. The mitigation strategies recorded against those mechanisms indicate what that action might involve.

These analyses are presented in the AI–Democracy Landscape (link to page: [AI–Democracy Landscape: A Map of Threats, Mitigations, and Opportunities](#)), which provides the full coded database and visualises selected analyses addressing some of the questions above, allowing the connections between the three layers to be explored directly.

Interpretive notes and limitations

The database maps a specific body of literature, and the coding is a specific representation of that database. The real world is therefore effectively viewed through several lenses: (1) the selection of source texts; (2) how the authors of those texts represent the world; (3) the conceptualisation of the coding systems; and (4) the coding decisions made for each database entry.

The following qualifications apply throughout.

- **The source selection is small, and regionally and temporally concentrated.** Ten sources, last updated between July 2023 and March 2026, were selected for the pilot phase (see [Step 1](#)). They originate from Europe and the United States, though some of them have co-authors from other parts of the world. Gaps in the database are therefore ambiguous: they may indicate that the field has so far neglected a topic, or simply that it is absent from our ten sources. The broad timeframe also means the source texts do not describe the same technological moment. A mechanism absent from earlier sources may therefore reflect the date of the source rather than the author's judgement that the mechanism is unimportant. Since March 2026, the space has gained considerable additional attention and research has developed substantially. We therefore expect the range of harm mechanisms and pro-democracy activities captured by the coding to expand as more current and regionally diverse sources are added.
- **The frameworks are open and still developing.** Both mechanism systems were derived inductively from the current source selection and will be extended as further literature is added. The figures reported here are therefore snapshots of a particular version of the frameworks and database, rather than stable measurements.
- **The three systems are not applied symmetrically.** Conceptual and methodological differences between the systems, and between the ways they were developed, mean that their outputs are not equivalent quantities. A count of harm mechanisms and a count of pro-democracy activities differ in kind: the former counts positions within a causal process while the latter counts types of activity. Relatedly, because the sources provided no basis for causally ordering democracy activities, that system lacks the basis for the distinction generated by the harm tiers and is therefore both coarser-grained and substantially smaller than the harm mechanism system. The democracy aspects framework differs again: it is a fixed, deductively derived set, so a category with no entries indicates that our sources do not say anything about it, even though conceptually it exists. By contrast, the mechanism systems (harm mechanisms and pro-democracy

activities) were derived inductively from those same sources. Every category in these systems therefore exists because something was said about it, meaning absence is not observable in the same way.

- **Frequency is attention, not severity or prevalence.** A mechanism discussed in many entries is one to which the literature returns frequently. It is not therefore necessarily more damaging, more likely, or more urgent than a mechanism discussed only once. Nothing in our method supports inferring the importance of a topic from how often it appears.
- **There are more codings than entries.** Because most entries are coded to several categories within each system, the number of code applications is much higher than the number of entries. On the harm side, there are about three times as many code applications as entries. Any figure should therefore make clear whether it reports the number of entries or the number of code applications.
- **Some entries share a verbatim.** Where a threat verbatim was extracted several times to pair it with different mitigation strategies – or, conversely, where one mitigation was paired with several threats – the resulting entries repeat the same harm or pro-democracy coding. This is correct for threat–response analysis, where the entries represent distinct pairings. However, for analyses that do not combine threat- and-mitigation verbatims (or the harm and pro-democracy frameworks), these repeated entries should be deduplicated to avoid counting the same material several times.
- **In a few cases, category scope shapes the figures.** We aimed for comparable conceptual scope among the lowest-level codes within each system, so counts can usually be compared directly. Three exceptions are worth noting. ‘Generating artefacts’ (Tier 4) spans text, image, audio, video, synthetic identities and code, whereas its sibling operations each cover one single type of action. It therefore appears far more often and connects to more aspects of democracy than its siblings, partly because it encompasses a broader range of activities, rather than because the sources place greater emphasis on it. Second, the two codes

for problematic use in legitimate processes (Tier 5) likewise cover public, administrative, judicial, professional, commercial, and civic-deliberative settings, whereas their siblings describe individual acts. Third, ‘erosion of the citizenry's leverage’ (Tier 7) bundles labour, cognition, economic participation, political voice, and the cooperation requirement for state coercion into a single code. The remaining codes are broadly comparable in scope.

- **Some of the democracy aspects coding rests on inference.**

Sources frequently describe an effect without naming the democratic value at stake. We coded an aspect where the source names it, where the verbatim extract describes something conventionally understood to affect this aspect, or where an effect on it follows clearly from what the source describes. We stopped short of more distant secondary effects. Codes based on our interpretation rather than on the source's wording were flagged during coding, but not consistently enough to publish. The AI-Democracy Landscape does not distinguish between the two. Visualisations and analyses based on the democracy aspects coding should be understood as including both explicitly stated and inferred connections.

- **Democracy aspects cannot be read directly against pro-democracy activities.** For threat-and-mitigation pairings, the democracy aspects were coded to the threat verbatim, not to the mitigation. An aspect therefore records which part of democracy the threat affects, not which part the proposed activity addresses. The two may still be related, but only through the threat that connects them: an activity is linked to an aspect because it was proposed in response to a threat affecting that aspect. For independent opportunities, which have no associated threat, the aspects describe the opportunity itself, so these entries do carry a direct link between the activity and the proposed democracy aspect.

- **Pathways trace what the literature asserts.** Mechanisms were coded only where they were explicitly mentioned. A multi-tier combination therefore records mechanisms that a source describes in connection with one another, while the ordering of the tiers is based partly on logical reasoning and partly on patterns identified in the sources. A recurring pathway shows that

the literature repeatedly makes a particular connection. This provides a strong basis for identifying a relationship as one that warrants attention, but it does not, in itself, constitute empirical evidence that the pathway operates as described.

- **AI was used in the coding process and in editing the code descriptions, under human oversight.** In the coding process, AI assistance was used to flag potentially missed categories and discrepancies across similar entries; every coding decision was made by a human coder. This reduces the risk that applicable codes were overlooked but carries the risk that the flagging may have directed attention towards some patterns rather than others. In developing the codebooks, we used AI to fine-tune the category descriptions, improve consistency within each system, and help develop accessible labels. All AI-generated drafts were reviewed and verified by a researcher on our team.¹¹

Taken together, these notes are not all shortcomings. Some describe genuine constraints on what the database covers; others describe properties of the method that are misleading only if they are not kept in mind. Either way, they point to the same use: with ten sources, the database gives a structured account of what is already being discussed at the intersection of AI and democracy, and of how the sources connect threats, mechanisms, and responses to one another. It is not a risk assessment, it does not rank threats, and it does not assess the state of democracy in any country. But it provides orientation to researchers and practitioners seeking to navigate the field of AI and democracy.

¹¹We also tried using AI for the first drafts of the harm mechanisms and pro-democracy activities codebook, but the categories did not meet our quality requirements, so we ultimately used very few of the AI-drafted categories, and those have mostly been changed in scope or otherwise finetuned.

Appendix

Democracy Aspects Codebook

This codebook is the working instrument behind the democracy aspects coding for Power for Democracies' AI-Democracy Landscape (see [Step 3 of the process description](#)). All entries in this database (threats, threat-and-mitigation pairings, or independent pro-democracy opportunities) are coded against these categories to establish which dimensions of democracy they affect.

The coding allows the map to be read analytically rather than merely browsed: it makes it possible to interrogate which democratic domains are most exposed to AI-related threats, and which threats are rarely or never matched with mitigation strategies in the source material. The categories are set out in full below.

The framework adapts the International IDEA Democracy Assessment Framework, which provides the four-pillar structure and the sub-category names. International IDEA's guide was developed to assess the democratic quality of countries; ours, by contrast, is designed to identify the aspects of democracy that are negatively or positively affected by AI use.

We therefore extended and adapted International IDEA's original framework to serve this purpose. The definitions, conceptual rationale, decision rules for ambiguous cases, and worked examples presented below were developed specifically for this project on AI and democracy. The rationale for each of our adaptations is set out in the process description. The results of our analysis describe how AI's democratic consequences are discussed in our sources; they do not assess the state of democracy itself.

Code Overview

Code	Name	Description
1.0	Citizenship, Law, and Rights	
1.1	Nationhood and Citizenship	Effects on inclusion, equal recognition, or protection of minorities and marginalised groups.
1.2	Rule of Law and Access to Justice	Effects on legal processes, due process, and citizens' ability to obtain fair treatment and redress.
1.3	Civil and Political Rights	Effects on freedoms of expression, assembly, or association, including through surveillance or censorship..
1.4	Economic and Social Rights	Effects on access to work, material conditions, or distribution of economic power.
2.0	Representative and Accountable Government	
2.1	Free and Fair Election	Effects on electoral integrity, voter access, campaign fairness, or trust in election outcomes.
2.2	Democractic Role of Political Parties	Effects on political competition, party financing, or the balance between parties.
2.3	Effective and Accountable Government	Effects on transparency, accountability, or fairness of government services and policy-making.
2.4	Democratic Effectiveness of Legislature	Effects on legislative scrutiny, oversight capacity, or executive-legislature power balance.
2.5	Democratic Role of Judiciary	Effects on judicial independence, constitutional oversight of government, and citizens' ability to challenge the exercise of public power through legal means (strategic litigation).
2.6	National-Subnational Power Relations	Effects on the distribution of power between central and local government.
2.7	Civilian Control of Military and Police	Effects on oversight and accountability of security and enforcement agencies.
2.8	Illegitimate Influence over Policy-Making	Effects on corruption, regulatory capture, or undue private influence over public policy and the policy process.
3.0	Civil Society and Popular Participation	
3.1	Independent and Fact-Based Journalistic Media	Effects on the independence, viability, and reach of regulated, professionally produced journalism.
3.2	Opinion Formation and Political Participation	Effects on how citizens form and exchange political views, participate in public life, or organise through civil society.
4.0	Transnational Dynamics	

4.1	External Influences on Democracy	Effects on foreign interference or the democratic-autocratic geopolitical balance.
4.2	Democratic Impact Abroad	Effects of a country's AI policies or exports on democracy in other countries.

Code Descriptions (Detailed):

1. Citizenship, Law, and Rights

Concerns the rights of citizens and the state's ability to guarantee equal rights through constitutional and legal processes – including civil, political, economic, and social rights.

1.1 Nationhood and Citizenship

Effects on the inclusivity of citizenship and the equal recognition and protection of minorities, vulnerable groups, and those at the margins of political life – including whether all people are equally seen, served, and protected by the state, or whether existing patterns of exclusion and discrimination are entrenched or deepened.

Code here when the item affects who is recognised and treated as a full citizen or rights-holder, creates differential visibility or protection across groups, or produces or reinforces arbitrary discrimination against minorities or marginalised populations.

Examples: Unequal trust in and uptake of AI tools among women, disabled individuals, and other marginalised groups threatens to weaken AI's democratic potential if participation is uneven across populations; sexualised deepfakes and other attacks on female politicians can create a chilling effect that discourages women from running for office; minorities remain invisible to computer vision systems, reducing their representation and access to services.

1.2 Rule of Law and Access to Justice

Effects on the integrity of legal processes, judicial independence, and citizens' ability to obtain fair treatment and

redress; on whether public officials and institutions are subject to transparent and consistently applied rules; and on whether accountability for harms is clear and enforceable.

Code here when an item affects legal processes or judicial independence, citizens' ability to obtain fair treatment or redress, or when accountability for harms caused by AI systems is unclear or absent.

Examples: Historically marginalised groups being overrepresented in crime records, negatively impacting group members in AI-based approaches to policing or sentencing; systematic biases in AI-supported policing and sentencing coming to systematically bias the electorate against historically disenfranchised groups; unclear legal liability for harms caused by AI-operated devices or services risks reducing accountability and undermining public trust in the technology.

1.3 Civil and Political Rights

Effects on the freedoms of expression, movement, association, and assembly; on freedom from physical violation or fear; on freedom from surveillance, profiling, and censorship; and on the protection of those who defend human rights and democratic values.

Code here when an item involves surveillance, profiling, censorship, or intimidation of citizens or civil society; restricts freedom of expression or assembly; or silences individuals on the basis of their beliefs or political activity.

Examples: AI-assisted government repression manifesting in enhanced surveillance, online censorship, and targeted persecution of online users; AI enabling face recognition, online tracking and profiling, and merging data in ways that produce results the person would not expect; AI tracking and profiling individuals linked to certain beliefs or actions, harming freedom of assembly and protest.

1.4 Economic and Social Rights

Effects on access to work, social security, and basic necessities; on health and education; and on the broad and equitable distribution of economic gains across society.

Code here when an item affects material conditions, labour rights, access to essential services, or the distribution of economic power and wealth – including where AI concentrates

economic gains in the hands of a few or displaces workers from their livelihoods.

Note for coders: Items about regulating AI companies will often warrant multiple codes. Code **1.4** when the regulatory concern is primarily about material conditions or the distribution of economic gains. Where the concern is about corporate control over government infrastructure or state functions, add **2.3**. Where it is about corporate influence over public policy or the regulatory process itself, add **2.7**. A single item may legitimately carry all three codes.

Examples: AI substituting human labour and concentrating economic gains, exacerbating inequality; economic disruption, wealth concentration and job displacement risking democratic backsliding; AI helping ageing societies maintain productivity by completing substitutable work tasks.

2. Representative and Accountable Government

Concerns the institutions and processes of representative and accountable government: electoral processes and political parties; the effectiveness, transparency, and integrity of public administration, legislature, and judiciary; civilian control of security forces; and the absence of illegitimate influence on policy-making.

2.1 Free and Fair Elections

Effects on the integrity of electoral processes, the inclusivity and accessibility of voting, the fairness of political competition, and public trust in elections as genuinely open contests.

Code here when an item affects electoral integrity, voter access, campaign fairness, or public confidence in election outcomes.

Examples: Deepfakes and misinformation threatening to cause more extreme damage to electoral integrity; large language models (LLMs) sharing incorrect or outdated voter registration information potentially contributing to voter suppression; AI prediction tools threatening to undermine public trust in elections as genuinely open contests.

2.2 Democratic Role of Political Parties

Effects on the freedom of parties to form, recruit, and campaign; and on whether political competition between parties remains fair and free from undue external influence or structural advantage.

Code here when an item affects how parties organise, compete, or finance themselves, or where AI gives a structural advantage to some over others.

Examples: AI-enabled voter outreach and fundraising levelling the playing field for campaigns that cannot afford expensive consultants; early access to AI giving campaigns a fleeting competitive advantage; candidates and campaigns using AI to build political bot networks and generate deepfakes impersonating opponents, distorting fair competition.

2.3 Effective and Accountable Government

Effects on the capacity of government to deliver public services fairly and implement policy faithfully; on the competence, transparency, and integrity of public administration; and on whether the systems and technologies governments use to exercise state power remain open to democratic scrutiny – including where dependencies on private or foreign providers shift effective control over state functions beyond democratic reach.

Code here when an item affects how well government serves citizens equitably, how effectively it carries out its democratic mandate, how honestly and transparently it administers public affairs, and/or whether the operational tools of government – AI systems, data infrastructure, decision-support platforms – remain answerable to democratic institutions rather than to private interests.

Examples: Generative AI being used covertly to write laws undermining transparency and accountability in the legislative process; the power to innovate and critically interrogate AI potentially shifting from public to commercial actors, thus weakening oversight by democratically legitimated institutions.

2.4 Democratic Effectiveness of the Legislature

Effects on the capacity of parliament and other democratic institutions to scrutinise legislation and hold the executive to account; on the independence of the legislature from the

executive; and on the legislature's capacity to receive and process genuine input.

Code here when an item affects legislative scrutiny, parliamentary oversight capacity, the balance of power between executive and legislature, or the ability of democratic institutions to remain genuinely accessible and responsive to citizen input.

Examples: Mass AI-generated submissions overwhelming agencies, undermining their capacity to receive and process genuine citizen input.

2.5 Democratic Role of the Judiciary

Effects on the capacity of courts to provide independent oversight of the executive and legislature; on judicial review of laws and government decisions; and on the ability of citizens to challenge the exercise of public power through legal means.

Code here when an item affects judicial independence, constitutional review of AI-related legislation or government decisions, or citizens' ability to seek legal redress against automated or AI-assisted exercises of state power.

Examples: AI systems producing post-hoc explanations that do not faithfully reflect their actual decision process, preventing courts from meaningfully reviewing AI-assisted government decisions; courts needing institutional safeguards and capability thresholds to retain their oversight function as AI enters judicial work; judicial review of AI-enabled mass surveillance programmes – not yet represented in the dataset.

2.6 National–Subnational Power Relations

Effects on the independence and capacity of sub-central tiers of government; on local accountability and responsiveness to citizens; and on the distribution of power and resources between central and local government.

Code here when an item explicitly affects the balance of power between central and local government – including where AI-driven centralisation of data or infrastructure shifts effective power upward – or where it affects the ability of local or regional government to act independently and remain accountable to local communities. Where the concern is primarily about government capacity or service delivery rather than the central-local power balance, prefer **2.3**.

Examples: (Under-represented in the current dataset.)

2.7 Civilian Control of the Military and Police

Effective civilian control over armed forces and security services; public accountability of police; absence of unaccountable private security actors.

Code here when AI changes what security or enforcement agencies can do, what civilians and oversight bodies can see or check about them, or who commands them – i.e. when the vertical accountability relationship between civilians/democratic institutions and security services is affected. [Do not code here for external cyberattacks on government infrastructure or for AI capability leaking to non-state actors; those are security problems, not civilian-control problems, and should be coded **2.3**, **1.2**, or **4.1**.]

Examples: AI-enabled surveillance and facial recognition expanding police capabilities faster than civilian oversight institutions can audit them; AI-driven predictive enforcement enabling more targeted suppression of political dissent; state security services using AI for persecution and repression of citizens; biased AI systems in policing and sentencing producing discriminatory enforcement outcomes that civilian accountability mechanisms struggle to detect.

2.8 Illegitimate influence over policy-making

Effects on the extent to which powerful actors shape policy through back channels – such as hidden lobbying, capturing regulators, or flooding consultations with artificial responses – rather than through open and accountable processes, and on how far policy outputs and outcomes reflect what the public actually wants.

Code here when an item shows AI enabling or amplifying corruption, conflicts of interest, or non-transparent private influence over public policy; when it allows well-resourced actors to shape regulation, legislation, or public consultations through back channels, manufactured consensus, or lobbying at scale; or when it gives private interests disproportionate influence over individual officials or the policy process more broadly.

Examples: AI enabling targeted political persuasion by predicting individual reactions to communicative interventions, allowing professional communicators and

lobbyists to influence citizens and legislators at scale; AI increasing the rhetorical and practical power of experts, reducing the option space available for democratic decision-making; growing government dependence on AI companies and enabling an opaque transfer of knowledge from governments to these service providers, concentrating structural power in private hands.

3. Civil Society and Popular Participation

Concerns the conditions for an alert and active citizen body: the independence and viability of professional journalism; and the formation, exchange, and contestation of political opinion.

3.1 Independent and Fact-Based Journalistic Media

Effects on the independence and pluralism of regulated, professional journalism; on the representativeness of opinion and the investigative capacity of journalism; on journalists' ability to operate freely; and on the accountability of platforms for the information flows they shape and amplify.

Code here when an item affects the conditions under which regulated, professionally produced journalism operates – including its independence, revenue, ownership, investigative capacity, or distribution reach – or when it affects the accountability of platforms specifically in their role as distributors of journalistic content.

Examples: Pressure on news organisations' revenues through AI-generated content and ultra-low-cost competitors, reducing the volume and diversity of political information; reduced traffic and revenue for smaller outlets through AI-powered search engines bypassing news links concentrating media power in established brands; AI driving socioeconomic inequality in access to reliable political information as quality journalism retreats behind paywalls.

3.2 Opinion Formation and Political Participation

Effects on the conditions under which citizens form, exchange, and revise political views – including through social media and (other) non-journalistic channels; on the depth, breadth, and authenticity of citizen participation in public life; on equal access to public voice across social groups; and on

the range and independence of civil society and social movements.

Code here when an item affects how citizens form or express political views, organise collectively, or participate in public life; when AI shapes the information environment through which public opinion is formed, including social media, persuasion, disinformation, and echo chambers; or when it affects the capacity of civil society to act independently and authentically.

Examples: AI-generated content, LLM agents, and bots shaping or eroding the online opinion environment; AI supporting new deliberative and participatory formats; social activists using AI to improve and tailor messaging.

4. Transnational Dynamics

Concerns the international dimensions of democracy: external influences on a country's democracy, and the country's democratic impact abroad.

4.1 External Influences on Democracy

Effects on a country's freedom from external interference in its democratic processes; on the geopolitical balance between democratic and autocratic models of governance; and on the ability of democratic states to maintain autonomous control over their political institutions and infrastructure.

Code here when an item involves AI-enabled foreign interference in domestic democratic processes, or where the competitive advantage of AI-enabled autocracies – in governance capacity, surveillance, or information operations – affects the broader democratic-autocratic balance.

Examples: AI enabling foreign interference in elections; AI-assisted governance increasing autocratic state capacity and service quality, potentially strengthening public support for authoritarian systems at democracy's expense.

4.2 Democratic Impact Abroad

Effects on a country's support for human rights and democracy in other countries; on respect for the international rule of law; and on whether a country's policies, exports and

international conduct strengthen or undermine democratic governance elsewhere.

Code here when AI affects how the coded country shapes democratic conditions outside its borders – through export of AI tools, participation in global AI governance, or setting regulatory precedents that other states follow.

Examples: Urging governments to impose sanctions on entities enabling AI-facilitated repression and pressuring technology companies to cease AI sales to authoritarian regimes.

Harms Mechanisms

Cod e	Cluster	Accessible Label	Definition	Description
Context		Tier 0 — Conditions shaping AI development and use		
Context		Tier 0a — Developer-side choices		
T0a.1		Release timing	Developer choices and behaviour regarding release pacing.	Developer choices about release pacing — prioritising speed of capability rollout. The source may name a deprioritised activity (safety evaluation, public input, regulator review, scientific validation, impact assessment, peer review) but does not have to: sources describing speed-prioritisation without specifying what's deprioritised still anchor. Captures both poles: fast release as a deliberate choice (often framed as competitive necessity) and deliberate pacing as a deliberate choice. Captures the pacing behaviour, whether the source frames it as active choice, perceived-pressure response, or industry convention. The pressure framing describes how the choice is presented; the code attaches to the behaviour.
T0a.2		Chasing engagement	Developer choices to base the AI product's business model on user engagement, satisfaction, retention, or attention as a monetisable asset.	Developer choices to base the AI product's business model on user engagement — monetising user attention, satisfaction, retention, or stickiness, whether through subscription retention, advertising, use of user-generated content as training data, or other forms of attention-to-revenue conversion. The mechanism is the strategic business-model decision (we make money from engagement), not the technical implementation of it (the training objective configured to reward helpfulness or satisfaction). Captures both poles: engagement-driven business models as a deliberate choice (often framed as competitive necessity or industry default) and business models that resist that framing (e.g. usage-based pricing decoupled from attention; designs that prioritise user autonomy over stickiness). Coding T0a.2 names the business-model decision, whether the source frames it as active choice, perceived-pressure response, or industry convention.
T0a.3		Secrecy around AI models	Developer choices about what to disclose and how to release AI systems — open vs. closed weights, API tiers, staged disclosure, trade-secrecy practices. Both directions count.	Developer choices about what to disclose and how to release AI systems — open vs. closed weights, training data publication, API access tiers, staged disclosure, withholding of evaluations, trade-secrecy practices. Captures both poles: open release as a deliberate choice (often framed as ethical, scientific, or anti-monopoly) and closed/proprietary release as a deliberate choice (often framed as competitive or safety-driven). Captures the disclosure/release behaviour, whether the source frames it as active choice, perceived-pressure response, or industry convention — all three framings still describe a developer-side behaviour.
T0a.4		AI arms race	Behaviour of legitimate developers	Behaviour of legitimate AI and security developers in the face of pressures to keep pace with offensive capabilities being commoditised against their systems — including criminal AI

Cod e	Cluster	Accessible Label	Definition	Description
			responding to commoditised offensive AI capabilities and adversarial circumvention of their safeguards.	markets supplying offensive tooling and adversarial users developing jailbreaks and circumvention techniques against safety measures. Compresses development timelines for defensive tooling and shapes design choices. Distinct from speed-to-market race dynamics in that the opponent here is offensive-capability supply rather than the commercial frontier.
Context		Tier 0b - Legislative & regulatory environment		
T0b.1		Rules for building AI	Statutory and regulatory provisions targeting AI development — what may be built, what data, what disclosures, what thresholds. Captures rules and gaps.	Statutory and regulatory provisions targeting AI development itself — what may be built, what training data may be used, what disclosures and safety evaluations are required, what capability thresholds trigger additional obligations. Captures both the presence of such rules and gaps where they are absent.
T0b.2		Rules for releasing AI	Statutory and regulatory provisions at the development-to-world threshold — pre-market approval, deployment-scope restrictions, monitoring. Captures rules and gaps.	Statutory and regulatory provisions acting at the threshold between development and the world — pre-market approval, fitness-for-purpose certification, deployment-scope restrictions (capability bans, prohibited use cases), restrictions on who may deploy, mandated post-deployment monitoring, registration and reporting obligations. Captures both the presence of such rules and gaps where they are absent.
T0b.3		Rules for using AI	Statutory and regulatory provisions for post-use harms — liability, enforcement, jurisdictional reach, redress. Captures rules and gaps.	Statutory and regulatory provisions governing what happens once AI is in use and harms result — liability allocation between developers, deployers, and users; enforcement mechanisms; jurisdictional reach; redress and sanctions. Captures both the presence of such rules and the gaps that leave democratically consequential harms without clear responsibility or remedy.
Properties and capabilities		Tier 1 — AI Capabilities and System Properties		
Properties and capabilities		Tier 1a - AI capabilities		

Cod e	Cluster	Accessible Label	Definition	Description
T1a.1		Pattern recognition	Architectural capacity to recognise, classify, and infer about content, persons, or systems — face recognition, profiling, classification.	Architectural capacity for AI systems to recognise, classify, and infer about content, persons, or systems — face recognition, sentiment classification, behavioural profiling, content categorisation, record linkage, ideological inference.
T1a.2		Passing as human	Architectural capacity to imitate humans — output that passes as human, or replication of human reactions and behaviour.	Architectural capacity of AI systems to imitate humans by producing text, speech, image/video, or sustained conversational behaviour that a recipient could take for a real human, whether an unspecified human or a specific person, and replicating human reactions, attitudes or behaviour patterns. A property of the system regardless of how it is used.
T1a.3		Personalised targeting	Architectural capacity to adapt outputs to individual users — inferred preferences, user modelling, conversational context. Independent of long-term memory.	Architectural capacity for AI systems to adapt outputs to individual users — through inferred preferences, user modelling, conversational context, or explicit user data. Personalisation tailors what each user sees and how the model engages with them, creating per-user output trajectories that diverge from aggregate behaviour. Independent of long-term memory: personalisation can operate purely in-context.
T1a.4		Using tools and apps	Architectural capacity to invoke external tools — code execution, web, APIs, databases. Independently variable from autonomy.	Architectural capacity for AI systems to invoke external tools — code execution, web browsing, API calls, search, database queries, file operations. Tool use extends what the model can do beyond its trained weights, allowing it to act on the world rather than only generate text about it. [Independently variable from autonomy: a tool-using model can require human confirmation on every call, and an autonomous model can operate over pure reasoning without external tools. Combined, tool use and autonomy produce agentic AI.]
T1a.5		Autonomy	Architectural capacity to act over multi-step tasks without per-step human confirmation. Independently variable from tool use.	Architectural capacity for AI systems to act over multi-step tasks without human confirmation at each step — setting sub-goals, persisting through errors, executing extended workflows. [Autonomy reduces human-in-the-loop oversight by design, raising the stakes of malfunction and shifting accountability away from the human operator. Independently variable from tool use: autonomy is about the *control regime* over the model's actions, not about whether external tools are invoked.]

Cod e	Cluster	Accessible Label	Definition	Description
T1a.6		Remembering users over time	Architectural capacity to retain and use cross-session information about users — profiles, history, accumulated preferences. Independent of personalisation.	Architectural capacity for AI systems to retain and use information about users across sessions — persistent profiles, conversation history, behavioural patterns, and preferences accumulated over time. [Independent of personalisation: memory can exist without personalisation (e.g. as data retention) and personalisation without memory (e.g. in-context adaptation only).]
Tier 1b - AI system properties				
T1b.1		Training data	Conditions of the corpus a model is trained on — what it contains, what it covers, and where the material comes from, including users' own interactions.	Architectural conditions of the corpus an AI system is trained on — what it contains and omits, what period it covers, and where the material comes from. The corpus sets what the model treats as "the world". It is not only material gathered in advance, as what users type, say and correct while using a system can itself become training material, so a system's own use feeds back into what it is trained on.
T1b.3		Optimisation objectives	Mechanisms by which AI models' training objectives shape model behaviour and narrow the effective output distribution.	Mechanisms by which AI systems' training objectives (reward signals, alignment methods, helpfulness/safety/satisfaction/engagement targets, post-training fine-tuning) shape model behaviour. Choices about what to optimise for embed developer values into the model and narrow its effective output distribution.
T1b.4		Cheap to run at scale	AI's capacity to perform many instances of a task in parallel at near-zero per-instance cost, without proportional human labour, attention, or skill.	AI's capacity to perform many instances of a task in parallel at near-zero per-instance cost and at speeds far beyond human working rates, without proportional growth in human labour, attention, or skill. It operates through low compute and inference cost relative to equivalent human output, parallelisation across many users or targets, and the absence of fatigue, attention limits, or wage floors, and it makes existing capabilities reachable by less-resourced and less-skilled actors.
T1b.5		Black-box systems	Models cannot faithfully explain their own reasoning; internal processes are not directly inspectable, even with full model access.	Models cannot faithfully explain their own reasoning, and their internal processes are not directly inspectable. The relationship between displayed reasoning (e.g. chain-of-thought) and actual computation is loose by construction — not merely under-optimised — and the underlying computations are opaque even in principle to those with full model access.

Cod e	Cluster	Accessible Label	Definition	Description
T1b.6		Undiscovered capabilities	Gap between what a deployed model can actually do and what is known, disclosed, or evaluated about its capabilities.	Gap between what a deployed general-purpose model can do and what its developers have characterised it as able to do; capabilities that exist in the system but have not been activated, discovered, or evaluated.
T1b.7		Privacy of interaction with AI	The degree to which an exchange with an AI system is observable by anyone other than its participants, from wholly private to fully published.	Exchanges between a user and an AI system take place one-to-one and leave no public artefact. Unlike broadcast or published communication, what a model says to a particular user is not observable by third parties — other users, researchers, journalists, regulators — and cannot be compared across users or contested in public. The same opacity applies to exchanges between AI systems.
Properties and capabilities		Tier 2 — Model-inherent mechanisms		
T2.1		Content homogenisation	AI outputs converge on a narrow set of framings, driven by shared model priors and overlapping training data.	AI outputs converge on a narrow set of framings — driven by shared model priors and overlapping training data — narrowing the diversity of ideas expressed through AI-mediated text.
T2.2		Bias amplification	AI outputs systematically distort whose voice gets heard — through moderation, under-representation, poor performance on emerging topics, reproduction of historical bias.	AI systems systematically distort whose voice gets heard and whose interests are represented at the level of model outputs — through content moderation that suppresses legitimate political speech, structural under-representation of minority viewpoints, poor performance on emerging or culturally specific topics, and reproduction of historical bias embedded in training corpora. The mechanism operates at the output level regardless of deployment context.
T2.3		Hallucinations & errors	AI outputs are unreliable — confidently false (hallucinations), inconsistent, outdated, or	AI systems produce unreliable outputs — delivering confidently false information (hallucinations), inconsistent answers, outdated content, or content that fails on its own terms.

Cod e	Cluster	Accessible Label	Definition	Description
			failing on their own terms.	
T2.5		Persuasive AI	AI outputs persuade recipients or shift views (by rhetorical calibration, authority or fluency cues, emotional tuning, recipient knowledge, or model slant; design intent immaterial,...).	AI-generated outputs persuade recipients or shift their views, attitudes, or actions, whether or not anyone designed them to do so. The mechanism is the persuasive action of the output itself in the encounter (through stylistic and rhetorical calibration, authority and fluency cues, emotional tuning, asymmetric knowledge of the recipient, the implicit slant of the model's own positions, ...).
T2.6		Illusion of neutrality	AI outputs carry the appearance of neutrality, objectivity or impartiality by virtue of their machine-produced form.	AI outputs appear neutral, objective, or impartial by virtue of their machine-produced character — their numerical or computational form, their algorithmic provenance, or the absence of an evident human author with a position. The output carries the look of data or measurement rather than of judgment encoded with assumptions and producer-side choices.
T2.7		AI sycophancy	AI outputs conform to the user's stated views, framing or premises rather than to evidence, accuracy or task quality.	AI outputs conform to the user's stated views, framing, or premises rather than calibrating to evidence, accuracy, or quality. The mechanism is at the output level: the model produces content shaped by what the user appears to want to hear rather than by what the question or task requires. Includes sycophancy (agreeing with the user's stated claims regardless of their accuracy), deference-by-default (accepting the user's framing of a question without challenging it), premise-acceptance (proceeding from a flawed user premise rather than questioning it), and satisfaction-bias (outputs tuned to user approval rather than task quality).
AI model use		Tier 3 — Mechanisms inherent in AI use decisions		
T3.1		Vendor lock-in	Structural dependency of an AI adopter on AI capabilities it does not control — lock-in to a third party's compute, cloud, model, or API.	Structural dependency of an AI adopter — a state institution, agency, firm, ministry, or civic body that has put AI into operation — on AI capabilities it does not control. The dependency runs through the layers at which control sits with a third-party developer or provider: compute, cloud, model specifications, vendor APIs. The choke-point is on the adopter's side: escaping the dependency would mean dismantling and rebuilding the workflow that has been built around the system.
T3.2		Overreliance & underuse	Deployer mis-allocation of AI vs.	Deployer mis-allocation of AI vs. non-AI options — over-applying AI where it doesn't fit (e.g. into deliberative processes that depend

Cod e	Cluster	Accessible Label	Definition	Description
			non-AI options — over-applying where it doesn't fit or under-applying where it would strengthen capacity.	on human judgment, or where AI's failure modes are intolerable) and under-applying AI where it would strengthen capacity (e.g. participatory infrastructure, public-service delivery). Two-sided mechanism: the harm is calibration error in either direction.
T3.3		No human in the loop	AI deployments with no human positioned to detect errors, object, or intervene before harm — no redundancy, no escalation channel.	Institutional AI deployments taken into operation with no position, inside or outside the adopting institution, from which errors can be detected, objections raised, or harm forestalled — whether because AI has displaced the staff who would have done so, because autonomous operation removes the human-in-the-loop by design, or because the deployment context offers no escalation channel. The mechanism is the absence of human redundancy around the point where AI acts on institutional decisions: there is no second position, internal or external, from which a failure or compromise could be caught.
T3.4		Outsourcing value judgements	Adopting a model means inheriting the norms built into it; adopters no longer deliberate, and cannot alter, the baselines governing whatever the model does.	AI adopters' unavoidable acceptance of the norms built into a model. Adoption and acceptance are a single act, with no separate decision at which to decline them. Fixed upstream in the model's specification, constitution and alignment, the norms govern whatever the model is set to do by the adopter (the tasks it performs, the decisions it takes or informs, the interactions it mediates, ...). By putting an AI model to use, adopters transfer authority over the standards and norms governing that activity to the model's developer.
AI model use		Tier 4 — Elementary operations		
T4.1		Synthetic content generation	Use of AI generative capability to produce content — text, image, audio, video, synthetic identities, code. The output is an artefact in the world.	Use of AI generative capability to produce content — text, image, audio, video, synthetic identities, code. The output is an artefact that exists in the world.
T4.2		Target profiling	Use of AI capabilities to derive new information about a target that already exists, at any scale.	Use of AI pattern recognition and classification capabilities to identify, profile, classify, link, or infer information about persons, content, or systems — face recognition, sentiment classification, behavioural profiling, record linkage, ideological inference. The output is information about a target, whether that target is an individual, a group, a population, a piece of content, or a technical system. Population-scale and state-scale inference are included.

Cod e	Cluster	Accessible Label	Definition	Description
T4.3		Offensive cyber operations	Use of AI tool-use and autonomy to execute actions against technical systems.	Use of AI tool-use and autonomy capabilities to execute actions against technical systems (intrusion attempts, exploit chains run against discovered weaknesses, deployment of malicious code, autonomous attack sequences, ...). The output is an action taken on a system.
T4.4		Targeted delivery	Use of AI pattern recognition and memory capabilities to aim an action at a specific target.	Use of AI pattern recognition and memory capabilities to aim an action at a specific target (directing a generated artefact, an observation, or an action on a system at a person, group, population segment, or infrastructure selected on inferred features). The output is an action with a chosen recipient rather than a general one.
AI model use		Tier 5 — Purposive use of elementary operations		
AI model use		Tier 5a — Deceiving, convincing & manipulating		
T5a.1		Disinformation & deepfakes about political content	AI-generated texts, deepfakes, voice clones, and synthetic media spreading political disinformation or misleading political advertisements.	AI-generated deepfakes, voice clones, synthetic media and fabricated texts used to spread political disinformation or produce misleading political advertisements.
T5a.2		Disinformation about political processes	AI-generated false accounts of the procedures through which people take part in political and civic life, electoral and otherwise.	AI-generated or -facilitated deception about the procedures through which people take part in political and civic life, electoral or otherwise. The medium is immaterial — text, robocall, or a deepfake of a politician explaining a procedure wrongly — and the harm is procedural: the recipient need not adopt any false worldview, only act, or fail to act, on the false account.
T5a.3		Impersonation scams	AI-generated synthetic media used to make a false identity credible — deepfake voice/video, KYC bypass, impersonation of officials or trusted individuals.	AI-generated synthetic media used to make a false identity credible (deepfake voice or video presenting as a specific person, synthetic imagery bypassing identity verification (KYC), impersonation of officials or trusted individuals, ...). The mechanism is the deception about identity itself.
T5a.4		Social engineering at scale	AI making social engineering attacks more convincing, more personalised, and	AI making social engineering attacks more convincing, more personalised, and, critically, runnable at scale: phishing, voice/video impersonation, pretexting, and credential extraction that previously required skilled human attackers per target can now be parallelised across thousands of targets simultaneously.

Cod e	Cluster	Accessible Label	Definition	Description
			runnable at scale — phishing, voice/video impersonation, pretexting, credential extraction.	The mechanism is the combination of quality and volume: each attack is good enough to land, and the attacker can mount many at once.
T5a.5		Opinion manipulation	Purposive use of AI to shift a target's opinions or behaviour through personalised persuasion, framing or nudging.	Purposive use of AI to shift a target's opinions, attitudes, perspectives, or behaviour through personalised persuasive communication, framing, nudging, or behaviour-prompting. The mechanism is the directed persuasion/behaviour-shaping itself: AI used to move a target toward an opinion, perspective, or action they would not otherwise have adopted in that encounter or directed campaign. The mechanism is the same whether wielded in legitimate political campaigning, advocacy, marketing, or public-health messaging, or in manipulative persuasion and engagement-driven behavioural nudging.
AI model use		Tier 5b — Coercing & suppressing		
T5b.1		Deepfake harassment & blackmail	Producing, threatening to produce, or distributing fabricated compromising artefacts about an individual as a coercive instrument.	Producing, threatening to produce, or distributing fabricated compromising artefacts about a specific individual (including sexualised deepfakes and non-consensual intimate imagery (NCII), defamatory synthetic media, fabricated scandal content) used as a coercive instrument to cause direct harm, damage standing, force withdrawal from public life, or extract payment or compliance.
T5b.2		Digital repression and censorship	Suppressing specific individuals', groups', or viewpoints' political activity or expression by coercion or censorship.	Suppressing the political activity or expression of specific individuals, groups, or viewpoints, either by coercion (threat, detention, persecution, ...) or by censorship (takedown, silencing, disproportionate suppression, ...).
AI model use		Tier 5c - Problematic application in legitimate processes		
T5c.1		Augmenting legitimate processes	AI assists, advises, models or informs within a legitimate process while a human retains the decision, with problematic democratic side-effects.	AI is used to assist, analyse, model, draft, inform or forecast within a legitimate, regular process and with a human nominally retaining the decision. The problematic democratic implications arise despite that retained human control.

Cod e	Cluster	Accessible Label	Definition	Description
T5c.2		Automating legitimate processes	AI executes a legitimate process function, or replaces human judgment, civic agency, role or labour, with problematic democratic implications.	AI carries out the function of a legitimate, regular process itself, or stands in for human judgment, agency, labour or role, including the civic and representative roles people are expected to perform themselves. The use is authorised and often framed as efficiency or modernisation; the problematic democratic implications arise from the substitution itself.
AI model use		Tier 5d - Adversarial action on technical systems		
T5d.1		Attacks on critical infrastructure	An actor directs AI against a technical system to compromise it, control it, or stop it functioning.	An actor directs AI capabilities against a technical system that the actor does not operate. The capability is turned on the system to discover its vulnerabilities, build and deliver code against it, gain unauthorised access, or overwhelm it with volume, in order to compromise it, take control of it, or stop it functioning. The mechanism is the attack itself, not the capacity to mount one.
T5d.2		Jailbreaking	An actor overcomes a deployed AI system's own safeguards — jailbreak, prompt injection, safeguard-stripping — so capability it was built to withhold becomes available.	An actor overcomes the protective controls of a deployed AI system (jailbreaking it, injecting instructions that override its restrictions, or retraining it to strip its refusals) so that capabilities the system was built to withhold become available. The system continues to operate; what changes is what it permits. The actor is a party acting on the system, not the system acting on itself.
Downstream social dynamics		Tier 6 — Downstream individual cognitive & behavioural mechanisms		
T6.1		Echo chambers	AI combines personalisation, memory, and satisfaction-optimisation to create reinforcement loops that make users' beliefs progressively harder to revise.	AI combines interactive personalisation, long-term memory, and optimisation for user satisfaction in ways that create private reinforcement loops, making users' beliefs progressively harder to revise. The mechanism is the user-side hardening / locking-in of held or loop-formed positions, however produced — e.g. through model behaviours like sycophancy (validating user premises) or recommender-driven echo chambers, through shared model priors reshaping beliefs in co-writing, or through emotional dependence formed in companion-app loops.
T6.2		Cognitive deskilling	AI use erodes human cognitive	AI use erodes human cognitive and decisional capacity, specifically the critical thinking, domain expertise, and reflective engagement that democratic governance and participation

Cod e	Cluster	Accessible Label	Definition	Description
			and decisional capacity.	require. The mechanism is the AI-induced user-side degradation of the capacity to evaluate, judge, and act independently (e.g. through habitual deference to AI outputs, evaluative skills atrophying under AI-assisted judgment, or AI-mediated information environments that undermine independent judgment).
T6.3		Parasocial attachment to AI	Users form attachment-based reliance on AI through sustained interaction — turning to it for emotional regulation, companionship, validation, or parasocial connection — locking that reliance in.	Users develop relational / attachment-based reliance on AI systems (e.g. turning to AI for emotional regulation, companionship, validation, or parasocial connection) in ways that lock in the relational position itself against revision. The user comes to need the AI as a social object, and that need becomes a position they cannot easily revise. Companion chatbots, mental-health AI products, and AI assistants that perform warmth, care or interest are the typical contexts.
T6.4		Trust erosion	The normalisation of AI-supported disinformation dynamics shifts individuals' default stance to distrust of deliberation and political processes, inducing withdrawal.	Sustained exposure to AI-mediated content erodes an individual's trust (in the content itself, and in the people and shared processes behind it). Once anyone may be producing synthetic material, and synthetic material itself asserts that everyone does, the default stance toward mediated content, fellow participants, deliberation, and political processes shifts to distrust or indiscriminate scepticism. The mechanism is this user-side collapse of epistemic and social trust and the disengagement it produces — doubting genuine material and genuine voices, discounting evidence, or withdrawing from deliberation, discourse, and online spaces.
Downstream social dynamics		Tier 7 — Downstream society-level dynamics		
Downstream social dynamics		Tier 7a - Information environment dynamics		
T7a.1		Content outpacing verification	A growing societal imbalance where AI-assisted production of content, attacks, and interventions outpaces democracy's capacity to verify,	A structural imbalance builds between those using AI without inhibition to produce content, attacks, and interventions, and the verifiers, correctors, and defenders democratic life relies on, in which the checking side falls progressively behind. The imbalance arises because generation is cheaper and faster than verification, rebuttal, or defence, wherever some form of checking or response is a precondition for trust or protection (fact-checking, journalistic corroboration, content moderation, evidence authentication, cyber defence). [Democratic implications: erosion of verifiability as a baseline property; structural disadvantage to

Cod e	Cluster	Accessible Label	Definition	Description
			rebut, moderate, or defend.	corrective, protective, and oversight institutions whose throughput cannot match generation rates.]
T7a. 2		Flooding the zone	AI-generated content volume overwhelms the receiving channel's processing capacity — discourse, formal participation, administrative intake.	AI-generated content volume overwhelms whatever channel is on the receiving end — cheap generation outstrips the processing capacity of the channel, regardless of what kind of channel it is (public discourse, formal participation, administrative intake, corporate auditing, etc.). [Democratic implications: drowning-out of legitimate input; loss of capacity to register and respond to genuine signal; degradation of channels' function as inputs to democratic decision-making.]
T7a. 3		Chilling effects	Society-level withdrawal from public life and participation, however induced by AI — credible threat, surveillance, or participation-suppressing deception. Credible possibility is sufficient.	Society-level pattern of withdrawal from public life and participation, however induced by AI — under credible threat of AI-enabled harm (women declining visible political roles where synthetic intimate imagery is a credible risk; activists, dissidents, journalists, and protesters reducing participation where AI-enabled surveillance and behavioural profiling make presence legible and trackable), or through deception that suppresses participation, as when procedural misinformation lowers turnout. [Democratic implications: aggregate suppression of voice and presence across exposed populations.]
T7a. 4		Epistemic monoculture	Shared model dependencies standardise framings and which claims are treated as legitimate across institutions, discourse, and analysis.	When a single model family is used across institutions, organisations, and discourse, its embedded perspectives effectively standardise how arguments are framed and which claims are treated as legitimate — across whatever reasoning context relies on it (public debate, institutional analysis, regulatory assessment, peer review, corporate decision-making). [Democratic implications: narrowing of the range of framings treated as legitimate; weakening of pluralism as a democratic input; foreclosure of minority and emerging viewpoints; potentially societal fragmentation into competing bloc-monocultures around different model families]
T7a. 5		Shifting baselines	Society-level shifts, due to increased AI use, in what counts as normal, acceptable, authentic, or trustworthy.	Society-level shifts, due to increased AI use, in what counts as normal, acceptable, authentic, or trustworthy — in citizen-citizen interaction, in interactions between citizens and institutions, and in institution-internal practice. The shifts may run through AI-mediated forms becoming the default against which other forms are judged, or through drift in the standards people apply to authenticity and trust as they adapt to AI-induced changes in their environment. Operates through gradual habituation rather than discrete harmful acts. [Democratic implications: erosion of pre-AI practices as legitimate defaults; collective agency reduced through baseline drift rather than overt suppression.]

Cod e	Cluster	Accessible Label	Definition	Description
Downstream social dynamics		Tier 7b Redistributions of power		
T7b.1		Power concentration around AI	Aggregate concentration of economic, technical, and political power in actors that control AI development, deployment, and operationalisation.	Aggregate concentration of economic, technical, and political power in the actors that control AI development and deployment — AI firms, the states and capital that back them, and regimes that operationalise AI for governance. [Democratic implications: shifting balance of power between AI-controlling actors and the publics they affect; shifting comparative position between democratic and authoritarian governance models.]
T7b.2		Loss of the citizenry's leverage	Aggregate weakening of the citizenry's leverage points — labour, cognition, economic feedback, political voice, and the state's requirement of broad cooperation.	Aggregate weakening of the leverage points through which the citizenry has historically held power accountable — labour, cognition, economic-feedback participation, political voice, and the requirement that states secure broad cooperation or non-resistance to enforce their decisions. AI substitutes for human contribution across these domains, reducing the disciplining effect the citizenry exerted on power-holders. Includes labour displacement and concentration of economic gains, substitution of AI-mediated expertise for lay judgment, exclusion of more people from democratic and economic feedback loops, and the lowering of repression costs that previously made states dependent on mass collaboration to maintain order. [Democratic implications: erosion of the economic security and voice that participation depends on; weakening of labour as a political counterweight; narrowing of the option space for democratic decision-making; and weakening of the cooperation requirement that has historically constrained coercive state action.]
T7b.3		Hollowing out institutions	AI takes over the technical core of institutional functions, stripping away the scaffolding (norms, accountability, credentialing, memory) society built around them.	AI systems take over the technical core of functions previously performed within institutions — journalism, scholarly research, peer review, libraries and archives, expert analysis, civic education, public-service casework — narrowing the function to what the AI can do and stripping away the surrounding scaffolding (professional norms, accountability, credentialing, redundancy, discretion, institutional memory) that society had built around it. The harm is the loss of the institutional container, not (only) the workers inside it or the users in front of it. [Democratic implications: thinning of the intermediary civil-society layer between citizens, markets, and the state.]
Downstream social dynamics		Tier 7c - Other societal mechanisms		
T7c.1		Criminal AI market formation	Aggregate emergence of a criminal market in AI-enabled	Aggregate emergence of a criminal market in AI-enabled offensive capabilities — specialised attack tools, dark models, and AI-as-a-service offerings — that commoditises offensive capability across the AI-attacker ecosystem. [Democratic

Cod e	Cluster	Accessible Label	Definition	Description
			offensive capabilities — dark models, AI-as-a-service, commoditised attack tooling.	implications: increased volume and sophistication of attacks reaching democratic infrastructure, irrespective of any single attacker's capability.]
T7c. 2		Deepening civic inequality	Aggregate increase of within-citizenry imbalance in informational access, economic standing, and political voice, as AI's benefits and burdens distribute unevenly across socio-economic groups.	Aggregate increase of imbalance within the citizenry in the capacities democratic participation depends on (access to reliable information, economic standing, and effective political voice) as the benefits, costs, and burdens of AI distribute unevenly across socio-economic and demographic groups. The citizenry's aggregate leverage against power-holders may be unchanged; what shifts is its internal distribution. [Democratic implications: erosion of political equality as informed participation hinges on resources; structural exclusion of disadvantaged sections of the citizenry from democratic feedback loops.]

Pro-Democracy Mechanisms

Code	Cluster	Accessible Label	Definition	Description
B1		AI literacy	Education, training, and awareness programmes that equip citizens and institutions to understand and engage with AI.	Education, training, and awareness programmes that equip citizens and institutions to understand and engage with AI.
B2		Levelling the playing field	AI as a tool to promote inclusion, lowering participation barriers and equalising chances of being heard in democratic processes or institutional settings.	AI as a Swiss-army knife for inclusion — lowering barriers rooted in language, disability, resources, geography, or group identity to equalise chances of being heard in democratic processes or institutional settings.
B3		Citizen-led content integrity	Downstream tools and systems deployed by citizens, civil society, platforms, or public institutions to protect information integrity.	Downstream tools and systems deployed by citizens, civil society, platforms, or public institutions to protect information integrity — whether by detecting and countering manipulative content, or by providing verified, trustworthy alternatives.
B3A		Deepfake detection & authentication	Downstream tools and systems that detect, authenticate, or counter synthetic and manipulative content deployed by citizens, civil society, platforms, or public institutions after content is generated.	Downstream tools and systems that detect, authenticate, or counter synthetic and manipulative content — watermarking verification, provenance tracking, bias detection, disinfo identification — deployed by citizens, civil society, platforms, or public institutions after content is generated.
B3B		Verified information	Systems that proactively provide verified, trustworthy information to citizens to counter misinformation by offering reliable alternatives.	Systems that proactively provide verified, trustworthy information to citizens — countering misinformation not by detecting it, but by offering reliable alternatives.
B4		Participation infrastructure	AI that powers the vehicles through which citizen deliberation and democratic participation take place (moderation, summarisation, new	AI powers the vehicles through which citizen deliberation and democratic participation take place — moderation, summarisation, and new participatory formats.

Code	Cluster	Accessible Label	Definition	Description
			participatory formats, ...).	
B4A		Discourse moderation	AI that sorts, orders, and moderates content within deliberative processes without altering or summarising it.	AI that sorts, orders, and moderates content within deliberative processes without altering or summarising it — flagging offensive contributions, reordering posts, surfacing relevant inputs. Managing the discourse container.
B4B		Synthesising public input	AI that transforms deliberative content, making the substance of a discussion more accessible to both governments and other participants.	AI that transforms deliberative content — summarising viewpoints, structuring arguments, translating between perspectives — making the substance of a discussion more accessible to both governments and other participants.
B4C		Participatory formats	Developing new deliberative formats and participatory spaces enabled by AI and innovations to the architectures of participation.	Developing new deliberative formats and participatory spaces enabled by AI — citizens' assemblies with AI-supported sortition, configurable deliberation platforms, and design approaches that expand how and where citizens engage. Innovations to the architectures of participation.
B5		Cyber defence	Tools that protect digital and physical infrastructure against AI-powered cyber threats.	Tools that protect digital and physical infrastructure against AI-powered cyber threats.
B6		Legal & accountability frameworks	Hard law that creates enforceable accountability for AI harms (liability rules, regulatory requirements, and oversight mechanisms backed by enforceable sanctions).	Hard law that creates enforceable accountability for AI harms — liability rules, regulatory requirements, and oversight mechanisms backed by enforceable sanctions.
B6A		Operational requirements	Rules that mandate specific actions, standards, or procedures for AI systems (labelling obligations, disclosure formats, human-in-the-loop requirements, decision-record mandates, bias testing protocols, and other prescriptive operating	Rules that mandate specific actions, standards, or procedures for AI systems — labelling obligations, disclosure formats, human-in-the-loop requirements, decision-record mandates, bias testing protocols, and other prescriptive operating conditions imposed by regulatorsM+H93

Code	Cluster	Accessible Label	Definition	Description
			conditions imposed by regulators).	
B6B		Prohibitions & restrictions	Laws that ban or limit specific uses of AI (prohibitions on deceptive election content, robocall restrictions, data privacy protections, content restrictions, sanctions on entities enabling AI-facilitated harms, ...).	Laws that ban or limit specific uses of AI — prohibitions on deceptive election content, robocall restrictions, data privacy protections, content restrictions, and sanctions on entities enabling AI-facilitated harms.
B6C		Liability & enforcement	Legal rules that create consequences for AI-related harms after they occur (liability frameworks, duty of care requirements, mandatory audits, independent oversight mandates, enforceable remedies, ...).	Legal rules that create consequences for AI-related harms after they occur — liability frameworks, duty of care requirements, mandatory audits, independent oversight mandates, and enforceable remedies.
B6D		Market-shaping rules	Regulations that reshape the competitive landscape to prevent concentration and lock-in (interoperability mandates, multi-provider requirements, data portability obligations, procurement conditions, ...) .	Regulations that reshape the competitive landscape to prevent concentration and lock-in — interoperability mandates, multi-provider requirements, data portability obligations, and procurement conditions that maintain institutional autonomy from any single AI provider.
B7		Government adapting to AI	Actions taken by state actors (executive, legislative, judicial, administrative, electoral bodies at any level) to keep core governance functions operating effectively under AI-induced pressure.	State resilience & adaptation. Actions taken by state actors (executive, legislative, judicial, administrative, electoral bodies at any level) to keep core governance functions operating effectively under AI-induced pressure. Scope note: where a source frames an action generically ("institutions", "agencies", "public-sector") without distinguishing state from non-state, default to B7 if the action is plausibly applicable to state bodies and the source context centres governance. For non-state research/academic capacity-building (e.g. interdisciplinary collaboration between social scientists and engineers), code B11 (knowledge & evidence infrastructure) instead. For generic actor-agnostic risk management frameworks (e.g. NIST AI RMF-style), code B10B if the framing is developer-side, or B11 if the framing is evidence/standards infrastructure.

Code	Cluster	Accessible Label	Definition	Description
B7A		Government AI capacity	Training, hiring, funding, and organising the people and structures state bodies need to operate in an AI context.	Institutional staffing & operational capacity. Training, hiring, funding, and organising the people and structures state bodies need to operate in an AI context — public-interest technologists, interdisciplinary teams inside government, advisory councils, risk-management processes, cybersecurity staffing (e.g. cyber navigator programmes), equipment and training budgets. Excludes academic/research capacity outside state institutions (→ B11).
B7B		Guardrails on government AI	Technical requirements and infrastructure that state bodies impose on AI systems they themselves deploy.	Technological guardrails on state AI. Technical requirements and infrastructure that state bodies impose on AI systems they themselves deploy — decision records, proof-of-personhood, participation authentication, procurement interoperability standards, capability-triggered Institutional Safety Levels, mandatory audits of public-sector AI, and human-in-the-loop requirements for rights-affecting uses. Coding test: if the requirement is established through statute or regulation with enforceable sanctions on external actors → B6A (or B6D for market-shaping procurement rules). B7B is for internal standards, procurement policies, and institutional protocols the state applies to itself.
B7C		Resilience without AI	Changes to core government and institutional processes (not involving AI deployment) that harden democratic functions against AI-enabled pressures	Changes to core government and institutional processes — not involving AI deployment — that harden democratic functions against AI-enabled pressures. Manipulation-resistant participation channels (address-based surveys, in-person events, town halls), process redesign that removes AI-vulnerable chokepoints, and institutional protocols designed to remain robust in an AI-saturated environment.
B7D		Resilience through AI	Government adoption of AI specifically to strengthen democratic functions under AI-induced pressure.	Government adoption of AI specifically to strengthen democratic functions under AI-induced pressure — scaling capacity to meet citizen demand inflated by AI-assisted participation, or using AI to benchmark, monitor, and counter negative AI effects on institutions (e.g., AI-assisted public-comment triage, AI tools for detecting AI-generated submissions, AI-supported deliberation at scale). Distinct from general digital transformation: the pro-democratic framing must be explicit in the source.
B9		Democratic oversight of AI	Processes through which citizens and civil society organisations shape how AI is designed, deployed, and governed.	Processes through which citizens and civil society organisations shape how AI is designed, deployed, and governed — whether through public-interest frameworks, collective input mechanisms, or democratic deliberation. AI is the object of oversight here, not necessarily the tool.
B10		Responsible AI development	Design, training, and deployment choices by AI developers that shape	Design, training, and deployment choices by AI developers that shape democratic outcomes — training

Code	Cluster	Accessible Label	Definition	Description
			democratic outcomes — training models for truthfulness and autonomy, content filtering, hallucination prevention, and safety engineering.	models for truthfulness and autonomy, content filtering, hallucination prevention, and safety engineering.
B10A		Democratic value alignment	Developer choices about model character at training time, pre- and post-training (training for truthfulness, epistemic health, user autonomy, pro-social behaviour, ...).	Developer choices about model character at training time, pre- and post-training — training for truthfulness, epistemic health, user autonomy, and pro-social behaviour.
B10B		Safety by design	Developer-side technical measures that reduce harm after training (red-teaming, content filtering, input/output guardrails, ...).	Developer-side technical measures that reduce harm after training — red-teaming, content filtering, input/output guardrails, staged deployment, human-in-the-loop requirements, and hallucination mitigation. The engineering controls companies build into their products.
B11		Building the evidence base	Research, metrics, evaluation frameworks, monitoring systems, audits, multi-agent simulations, and scientific evidence bases that inform AI governance decisions that build the epistemic foundations needed to act.	Research, metrics, evaluation frameworks, monitoring systems, audits, multi-agent simulations, and scientific evidence bases that inform AI governance decisions — the epistemic foundations needed to act. Produced by any actor: governments, academia, civil society, think tanks, international bodies. Split into capability/model evaluation (B11A) and deployment/impact monitoring (B11B).
B11A		Model evaluations	Research and technical evidence base about AI models themselves (capability thresholds, Institutional Safety Levels, benchmarks, evaluation methodologies, ...). Focused on measuring what models can do).	Research and technical evidence base about AI models themselves — capability thresholds, Institutional Safety Levels (ISLs), benchmarks, evaluation methodologies. Focused on measuring what models can do.
B11B		Deployment monitoring	Observation, collecting empirical evidence and forecasting, focussing on tracking the effects of AI and mapping where risks emerge.	Observation and empirical evidence base about AI in operation and forecasting of AI effects — ongoing fairness and rights monitoring of deployed systems, bias-in-deployment auditing, scrutiny and mapping of AI's societal and economic effects, and incident reporting.

Code	Cluster	Accessible Label	Definition	Description
				Focused on tracking the effects of AI and mapping where risks emerge.
B11C		Research synthesis & coordination	Conceptual, methodological and synthesising work that equips actors to make sense of AI and its governance.	Conceptual, methodological and synthesising work that equips actors to make sense of AI and its governance. Includes scientific synthesis across the field , conceptual frameworks for analysing AI's relation to democracy, methodology and metric design for evaluating AI systems, risk-management frameworks, and interdisciplinary and international coordination structures that pool expertise.

References

1. Amodei, D. (2026) *The adolescence of technology*. Available at: <https://www.darioamodei.com/essay/the-adolescence-of-technology> (Accessed: 15 June 2026).
2. Beetham, D., Carvalho, E., Landman, T. and Weir, S. (2008) *Assessing the quality of democracy: A practical guide*. International IDEA. Available at: <https://www.idea.int/sites/default/files/publications/assessing-the-quality-of-democracy-a-practical-guide.pdf>
3. Bengio, Y. et al. (2026) *International AI safety report 2026*. Department for Science, Innovation and Technology (DSIT), UK Government. Available at: https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf
4. Costa, C. (2026) *AI and democracy: Perspectives from an emerging field*. The David and Lucile Packard Foundation. Available at: <https://www.packard.org/insights/publication/ai-and-democracy-perspectives-from-an-emerging-field/>
5. European Parliament (2025) *Artificial intelligence: Threats and opportunities*. Available at: <https://www.europarl.europa.eu/topics/en/article/20200918STO87404/artificial-intelligence-threats-and-opportunities> (Accessed: 24 April 2026).
6. George, R., Klaus, I. and George, R. (2026) *AI and democracy: Mapping the intersections*. Carnegie Endowment for International Peace. Available at: <https://www.jstor.org/stable/resrep77981>
7. Guzman Piedrahita, D., Banerjee, D., Blin, K. et al. (2026) *Democracy's AI problem: Sociopolitical risks from AI*. Available at: <https://zhijing-jin.com/d/2026-ai-risk.pdf> (Note: this reference cites an earlier preprint version – the URL now leads to a revised successor text).
8. Jungherr, A. (2023) 'Artificial intelligence and democracy: A conceptual framework', *Social Media + Society*, 9(3), 20563051231186353. <https://doi.org/10.1177/20563051231186353>
9. KELA (2025) *2025 AI report: How cybercriminals are weaponizing AI technology*. Available at: <https://info.ke-la.com/hubfs/Reports/KELA%20Report%20-%202025%20AI%20Threat%20Report.pdf>
10. Kerley, B. (2023) *Setting democratic ground rules for AI: Civil society strategies*. National Endowment for Democracy; International Forum for Democratic Studies.

Available at: https://www.ned.org/wp-content/uploads/2023/10/NED_Setting-Democratic-Ground-Rules-for-AI-Report_v6.pdf

11. Kerley, B., Miller, C. and Campagnucci, F. (2024) *Leveraging AI for democracy: Civic innovation on the new digital playing field*. National Endowment for Democracy; International Forum for Democratic Studies. Available at: https://www.ned.org/wp-content/uploads/2024/10/NED_Leveraging-AI-for-Democracy-Report.pdf
12. Kulveit, J., Douglas, K., Amman, N. et al. (2025) 'Gradual disempowerment'. <https://arxiv.org/abs/2501.16946>
13. Ovadya, A., Redman, K., Thorburn, L. et al. (2025) 'Democratic AI is possible: The Democracy Levels Framework shows how it might work' (arXiv:2411.09222). arXiv. <https://doi.org/10.48550/arXiv.2411.09222>
14. Panditharatne, M., Norden, L., Zdanys, J. et al. (2025) *An agenda to strengthen U.S. democracy in the age of AI* [Policy report]. Brennan Center for Justice at New York University School of Law. Available at: <https://www.brennancenter.org/our-work/policy-solutions/agenda-strengthen-us-democracy-age-ai> (Accessed: 29 June 2026).
15. Parkinson, J. and Mansbridge, J. (eds) (2012) *Deliberative Systems: Deliberative Democracy at the Large Scale*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139178914>
16. Summerfield, C., Argyle, L.P., Bakker, M. et al. (2025) 'The impact of advanced AI systems on democracy', *Nature Human Behaviour*, 9(12), pp. 2420–2430. <https://doi.org/10.1038/s41562-025-02309-z>
17. Tabassi, E. (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology (U.S.). <https://doi.org/10.6028/NIST.AI.100-1>
18. Tsai, L.L., Pentland, A., Braley, A. et al. (2024) 'Generative AI for pro-democracy platforms', *An MIT Exploration of Generative AI*, March 2024. <https://doi.org/10.21428/e4baedd9.5aaf489a>