SAFEGUARDING ELECTIONS IN THE AGE OF AI AND SYNTHETIC CONTENT: A FRAMEWORK FOR ELECTORAL INTEGRITY INSTITUTIONS

ALEŠ ČÁP
SIR GEOFF MULGAN

WHITE PAPER #001

FEBRUARY 2025

THE INSTITUTIONAL ARCHITECTURE LAB

WHAT IS TIAL

The Institutional Architecture Lab was formed in 2023 by Sir Geoff Mulgan, Jessica Seddon and Juha Leppänen in an effort to help the institutional design community coalesce, learn together, and grow. Each of us has been involved in various stages of creating new organisations and other institutions. Like many other people, we have witnessed first-hand the absence of a formal community along the way — or a place where we can learn from past experience. We are aware that there is a lot of great work happening around the world, but nowhere to recognise it.

AUTHORS

Aleš Čáp
UCL

Sir Geoff Mulgan
TIAL, UCL

Designed by Kilda

TABLE OF CONTENTS

EXECUTIVE SUMMARY

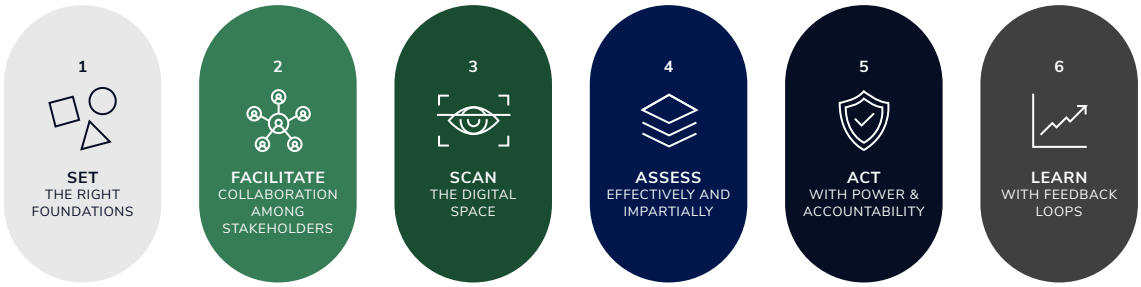
The rise of generative AI and synthetic content—such as deepfakes, manipulated audio, and AI-generated misinformation—has introduced unprecedented threats to electoral integrity, with the potential to erode public trust, deepen societal divisions, and destabilise democratic processes. While fears of deepfake-driven election disruption in 2024 were widespread, their actual impact was less decisive than many anticipated. However, **there is a risk of moving from undue alarm to unwarranted complacency**. As synthetic content tools become more advanced, accessible, and seamlessly embedded into digital ecosystems, their influence on elections is set to increase. Yet at present there are few institutions with sufficient capabilities or powers to respond.

This white paper makes the case for establishing **Electoral Integrity Institutions** to address these evolving threats.

It proposes a framework that offers a structured pathway that countries can adapt to their unique institutional, political, legal, and cultural contexts. It is not a rigid prescription but aims to expand the pool of viable options for countries that want to safeguard their democratic processes against abuses of synthetic content.

Drawing on real-world examples, academic evidence, cutting-edge methods, and lessons from successful (and less successful) institutions, this paper provides answers to some of the most pressing questions relevant to those designing an Electoral Integrity Institution in the era of synthetic content, showing how they can scan, assess and act decisively to protect elections.

The Framework: The white paper introduces a six-step framework for designing and operating future-ready Electoral Integrity Institutions (EII):



1. *SET the right foundations:*

   EIIs must begin with clarity on their task, people, and ethos. Institutions falter when these foundations are not clearly established or when their design and operations fail to reflect them.

2. *FACILITATE collaboration among stakeholders:*

   EIIs must **act as a hub for coordinating diverse stakeholders**, including government agencies, technology platforms, civil society, and academia. This paper outlines strategies and mechanisms for fostering collaboration, aligning incentives, and ensuring multistakeholder governance is effective.

3. *SCAN the digital space:*

   Proactive and comprehensive scanning is essential for detection of disinformation campaigns. EIIs should deploy a combination of advanced automated tools, a large network of volunteers, and expert analysts to monitor the digital landscape. By **seamlessly combining advanced technology with human judgement**, EIIs can rapidly detect and flag synthetic content, enabling timely intervention before false narratives gain traction.

4. *ASSESS content effectively and impartially:*

   Once content is identified, EIIs should undertake a rigorous triage process to evaluate its potential harm, degree of falseness, reach, virality, and source credibility. A tiered system is proposed that would determine the appropriate response. This process can ensure that **each threat is addressed proportionately, efficiently, and in alignment with democratic values** and legal standards.

5. *ACT with power and accountability:*

   To counter disinformation effectively, EIIs must be empowered to act decisively, and fast, while maintaining transparency and accountability. This paper proposes designing an interface with independent, democratically accountable powers—such as electoral commissions, courts, and multistakeholder boards. These measures must strike a delicate balance between protecting elections and upholding freedom of expression.

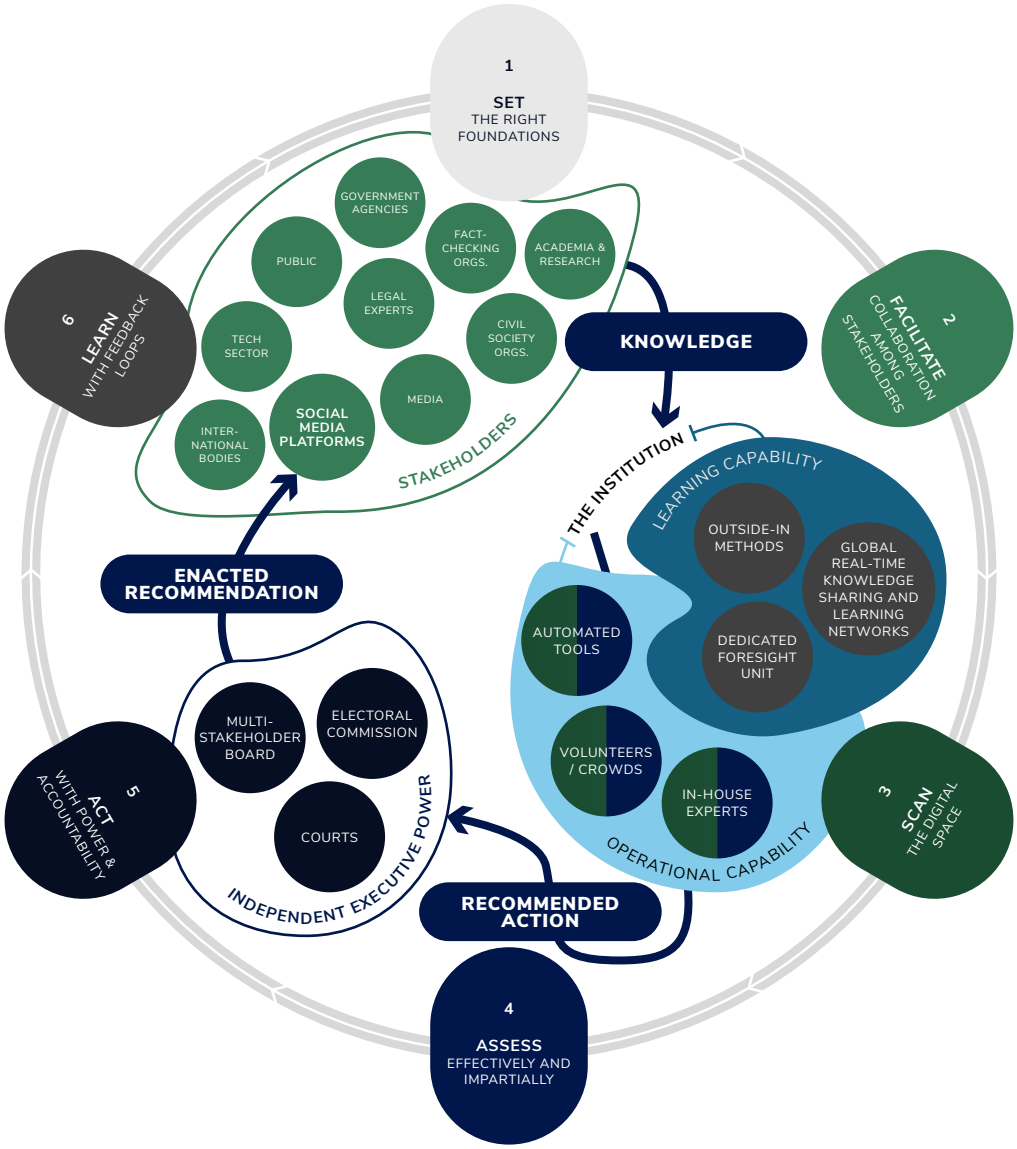
6. *LEARN via feedback loops:*

   In a constantly shifting digital environment, EIIs must learn and adapt continuously. By integrating lessons from its actions, engaging with global counterparts, incorporating external insights, and leveraging the expertise of its dedicated Foresight Unit, the institution can monitor and anticipate emerging threats. This **continuous feedback process** enables the institution to refine its strategies, ensuring it remains agile and resilient in combating synthetic disinformation.

The framework is grounded in **real-world case studies and global examples**, offering concrete lessons for institutional design. For instance, *Sweden's Psychological Defence Agency* demonstrates how a clearly defined mandate can effectively safeguard national security while ensuring that free speech rights are not undermined. *Taiwan's participatory policymaking* showcases innovative methods for engaging diverse stakeholders, fostering public trust and collaboration in tackling disinformation. *France's VIGINUM* illustrates the importance of in-house expertise and robust oversight mechanisms in protecting public discourse during elections. Conversely, the controversies surrounding the *Global Disinformation Index* highlight the critical need for mechanisms that ensure impartiality, alongside transparency, to maintain trust and credibility. These examples, among others, provide actionable insights and ensure the framework is both grounded in reality and informed by successes and failures worldwide.

Key Takeaway: This white paper provides a framework for countries seeking to establish Electoral Integrity Institutions, equipping them to tackle the complex challenges posed by synthetic content. Rather than prescribing a one-size-fits-all solution, it expands the range of viable options for institutional design. By implementing this framework, countries can safeguard their electoral processes and strengthen public trust in democratic institutions.

Figure 1. How the Electoral Integrity Institution operates: a high-level visualisation



WHO THIS PAPER IS FOR

This paper is primarily intended for national governments, policymakers, and anyone with a responsibility to address the challenges posed by synthetic content in the context of electoral integrity.

While the primary focus is on the national level, the insights and lessons presented here are also relevant to multinational organisations, international regulators, and other transnational bodies involved in addressing synthetic content and disinformation. For these stakeholders, the paper offers guidance on how national-level efforts can align with broader regional or global strategies.

While the focus is on synthetic media, the framework has broader applicability. Those working on disinformation in any form may find the insights presented here valuable for addressing the complexities of the modern information landscape.

Additionally, the paper hopefully can serve as a resource for academics, researchers, and civil society organisations interested in understanding the institutional dimensions of electoral integrity and exploring how diverse sectors can collaborate to counter the evolving threat of synthetic content.

By offering a structured and adaptable framework, this paper seeks to contribute to a wider dialogue on protecting democracy in an age of rapid technological disruption.

HOW TO READ THIS PAPER

This paper is structured to guide readers through the challenges posed by synthetic content and the proposed institutional framework for safeguarding electoral integrity. While all sections contribute to the overarching argument, readers may choose to focus on specific parts depending on their familiarity with the topic.

- *Section 1 (Introduction):* Provides a broad overview of synthetic content and its implications for elections.
- *Section 2 (Context):* Explores why an institutional response is necessary, examining the limitations of existing approaches and institutions–and the governance gaps they leave behind.
- *Section 3 (The Framework):* This is the heart of the paper. It outlines the proposed institutional designs in detail, offering actionable insights and strategies for implementation.

Throughout the paper, case studies and real-world examples are used extensively to illustrate key points. These examples, drawn from a variety of contexts and sectors, provide practical lessons and cautionary tales, highlighting both what works and what doesn't. From Sweden's Psychological Defence Agency to the cautionary tale of Global Disinformation Index, these cases bring the discussion to life, offering real-world insights for those designing or contributing to electoral integrity institutions.

INTRODUCTION

Artificial Intelligence (AI) is transforming much of daily life, the organisation of the economy and politics. The recent advent of powerful generative models has enabled machines to autonomously create content that mimics human-like patterns, revolutionising fields ranging from image synthesis to natural language processing. In conjunction with other GenAI related developments (e.g., LLMs), the coming wave of AI evolution features looks set to further amplify capacities to create synthetic content such as deepfakes.[1]

While synthetic content offers creative possibilities, its misuse threatens societal, political integrity, privacy, and national security, including micro-targeted assaults and election manipulation[2].

This paper focuses on elections and the institutional make-up necessary to protect the electoral integrity. It was drafted during 2024, dubbed the "Year of Democracy" when an unprecedented number of people worldwide headed to the polls, marking a significant surge in electoral participation and reflecting a global commitment to democratic engagement and the fundamental right to vote.

While early evidence[3] suggests that elections in 2024 may not have been severely disrupted by synthetic content—though this is difficult to confirm due to the limited access to social media platform data—the risks remain deeply concerning. The speed, scale, and sophistication of AI-generated content could erode public trust, deepen societal divisions, and destabilise electoral processes if left unaddressed.

Institutional adaptation is necessary to keep pace with the changing technological and informational landscape. The rapid advent of deepfake technology has created a largely unregulated environment, making it alarmingly easy to synthetically create or manipulate visual, audio, text information. This ease of creation and dissemination by virtually anyone underscores the urgent need for robust safeguards to protect the democratic process from technological interference.

SYNTHETIC CONTENT: MORAL PANIC OR A REAL THREAT?

'Synthetic Content' (Generated by Dall-E)

The technology underlying deepfakes relies heavily on Generative Adversarial Networks (GANs), which consist of two neural networks—a generator and a discriminator—working in adversarial collaboration. Advances in GAN architecture and training methodologies have significantly enhanced the realism of synthetic content, reducing the amount of data needed for training while leveraging increasingly expansive datasets. These improvements are driving deepfakes closer to a level of realism that renders them virtually indistinguishable from authentic video content[4].

The open-source nature of much of this progress, coupled with the availability of user-friendly software, has democratised the creation of deepfakes. Even individuals with minimal technical expertise can now produce convincing synthetic content. Experts predict that a substantial percentage of online material could soon be synthetically generated[5], a trend underscored by the recent rapid growth of fully AI-generated websites[6].

GenAI and its political applications have sparked concerns about sophisticated manipulation that could undermine democratic processes. While deepfakes can have positive or harmless uses, their misuse poses serious risks, such as discrediting political figures, spreading false information, and inciting social conflict[7]. The increasing accessibility of deepfake technology exacerbates these threats, as almost anyone can now produce convincing fake videos with minimal effort and cost[8].

As Logically reports[9], mass disinformation campaigns, like Russia's efforts to influence the 2016 US election, have traditionally been well-organised and well-funded. The Internet Research Agency (IRA), responsible for this operation, had around 400 employees earning up to 10 dollars per hour, with a 2017 budget of 12.2 million USD. This initiative posted over a thousand pieces of content weekly across 470 social media pages, reaching up to 126 million people on Facebook. It was a human-driven operation.

GenAI is set to dramatically lower the costs associated with such disinformation campaigns. Previously, large-scale disinformation required significant resources and skilled, organised teams, typically limiting such operations to state actors. GenAI can now automate much of the content creation process, producing culturally nuanced outputs with fewer telltale signs of inauthenticity.

This potential was recently highlighted by a 'proof of concept' system called *CounterCloud*. CounterCloud's AI identifies articles by specific publications and generates similar content using Large Language Models (LLMs). It also creates fake comments, images, and sound clips, and manages social media activity to promote or counter narratives. The end-to-end

process is automated, and the entire system was implemented for just 400 dollars. This shift means that the capacity to launch disinformation campaigns is no longer confined to state actors, as the barriers to entry are significantly reduced by AI-driven automation.

Finally, an emerging body of literature suggests that some of the most profound potential harms of synthetic content to elections may be less obvious, visible, and more incremental. Proliferation of synthetic content may lock individuals into their "personalised truths" and undermine our capability for *collective deliberation*[10]. Synthetic media, like deepfakes and AI-generated misinformation, can create echo chambers where people are continuously exposed to information that reinforces their existing beliefs. This fragmentation erodes the shared reality essential for democratic debate, leading to increased polarisation and hostility. As individuals become more entrenched in their micro-tailored narratives, the capacity for informed, consensus-driven decision-making diminishes, weakening the integrity of the electoral process, and its function as a democratic institution[11].

THE THREATS WE'VE SEEN

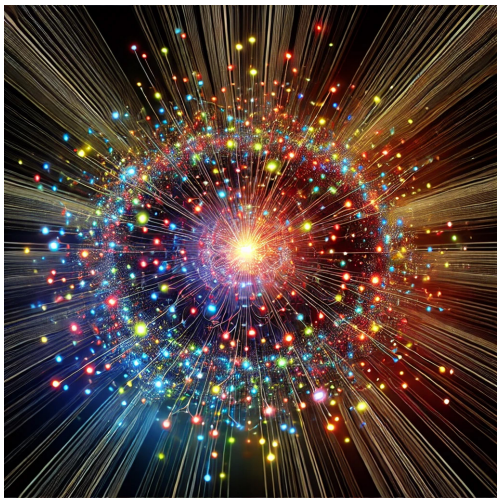
Recent electoral contexts have seen the use of synthetic content aimed at affecting the outcomes. During the US primaries, the widely-reported Joe Biden robocalls deepfake incident involved manipulated audio clips designed to mimic Joe Biden's voice. These robocalls falsely suggested that voting in the primary would preclude voters from later casting ballots in November. In Türkiye, deepfakes manipulated public opinion, forcing a candidate to withdraw. Argentina saw a storm of deepfakes used both for self-promotion and attacks during its presidential election[12]. In Slovakia's 2023 election, an audio deepfake surfaced two days before the vote, falsely depicting a scheme to rig the election[13]. This attack was insidious due to its timing and audio-only format, making it difficult to debunk. Ultimately, the true impact of this deepfake remains hard to measure, but many speculate that it might have played a decisive role and there is no doubt that the proliferation of synthetic content can breed distrust and uncertainty among voters, politicians, and journalists[14].

Recent evidence from The Alan Turing Institute[15] indicates that while its hard to measure the impact of synthetic content on election results, it's clear that AI-generated disinformation can amplify political polarisation by reinforcing pre-existing beliefs, as observed in the 2024 U.S. presidential election. As the Turing report suggests, this underscores the need for institutions to not only address the dissemination of false content but also to mitigate its societal impacts, such as deepening ideological divides.

The advent of GenAI thus adds yet another layer of complexity to an already precarious information landscape, plagued by misinformation, echo chambers, and polarising recommendation algorithms.

POSITIVE IMPACTS: FROM ONE-WAY BROADCASTING TO INTERACTIVE BROAD LISTENING

Deepfake technology is often seen through the lens of risk. But it can also be used in more positive ways, enhancing political communication and making politics more participatory, accessible, and even interactive. One recent example from India highlights this potential: Prime Minister Narendra Modi employed the government-backed AI tool Bhashini to translate his speech in real-time from Hindi to Tamil during a political address[16]. This showcased how AI and deepfake technologies can help bridge linguistic divides, allowing politicians to engage more effectively with diverse populations.



'Broad Listening' (Generated by Dall-E)

Looking ahead, deepfake technology could enhance political engagement by enabling candidates to create digital avatars. These avatars could present policies, hold debates, or interact with voters in a way that feels personal and immediate, without requiring physical presence. For instance, political candidates could use these tools to host virtual town halls, offering customised messages or real-time engagement with citizens. This aligns with the growing role of virtual spaces, such as the Metaverse, in hosting dynamic, interactive political campaigns.

Traditionally, political communication has been a one-way street, with politicians broadcasting messages to the public. However, forward-thinking leaders like Audrey Tang in Taiwan are pioneering a shift towards "broad listening" — a more participatory form of engagement where technology allows people to be involved in political decisions. In her co-authored book *Plurality*, Tang explains how platforms like vTaiwan enable citizens to collaboratively debate and co-create policies, leading to impactful results on issues ranging from Uber regulations to digital privacy.

This evolution of political engagement represents a significant opportunity for deepfake technologies to be used not just for broadcasting but for "broad listening." By creating interactive platforms where voters feel heard and engaged, politicians can foster two-way communication, enhancing the democratic process.

CONTEXT: WHY THIS REQUIRES AN INSTITUTIONAL SOLUTION

Democracy has always relied on an array of institutions, including not just parliaments and courts, but also institutions to manage elections or counter corruption.

Nobel Prize winner Douglass North defined institutions as "the humanly devised constraints that structure political, economic, and social interaction", and there is now widespread acceptance that these play a vital role in helping societies to thrive and prosper. Recent research emphasises their decisive role in supporting economic growth, equity and stability,[17] guiding nations along what

Nobel Prize winner Daron Acemoglu has called the "narrow corridor." The "narrow corridor" leads to peace and prosperity[18] often helping to distribute power broadly rather than concentrating it, fostering accountability, and preventing the abuse of power.

Every era requires a different mix of institutions. The advent of radio and television prompted much creative institutional design (such as the invention of the BBC in the UK or the FCC in the US).

As we show, the advent of the Internet, and more recently of AI, has also prompted institutional innovations ranging from ICANN to Sweden's Psychological Defense Agency and Moldova's Centre for Combating Disinformation. Existing institutions are struggling to respond adequately and reliance on social norms is also becoming problematic.

In the UK for instance, implicit norms have historically been a cornerstone of electoral integrity, with institutions like Ofcom and the Electoral Commission playing a relatively hands off role. The former mainly oversees the broadcasters and adverstising, and the latter concentrates on the financial aspects of elections, with clear accountability for oversight of the virtual space currently lacking–bar perhaps The Defending Democracy Taskforce, whose activity, however, is near-impossible to trace. With over-reliance on cultural norms comes a risk of opening up more space for malicious actors.

Many have argued that the changes in technological landscape have created incentives to 'cheat', and that cultural norms have been eroding. In the UK, this was evidenced by incidents like the Conservative Party temporarily rebranding their Twitter account to impersonate a fact-checking organisation during the 2019 election — changing their Twitter name to FactCheckUK and tweeting about a debate between Corbyn and Johnson. Importantly, the Conservative Party faced no significant repercussions for this action and experienced little reputational damage (and won the election in a landslide), highlighting the vulnerability of relying solely on cultural safeguards.

These examples and many others around the world point to the need for institutions to protect democracy, helping to create informational environments where trustworthy sources of information are easily accessible, and people are incentivised to seek the truth and not get stuck in their synthetic, personalised echo chambers[19].

WHAT'S OUT THERE? EXISTING ELECTORAL INTEGRITY INSTITUTIONS

In response to the increasing threats of foreign interference, disinformation, and other challenges posed to democratic processes, several countries have taken significant steps in establishing institutions to safeguard electoral integrity. These institutions are among the first movers in the domain, providing critical lessons for other countries looking to bolster their electoral systems.

The Electoral Integrity Assurance Taskforce (EIAT) in Australia, for example, was established to maintain public trust in the country's electoral processes. The Taskforce brings together various government agencies, including the Australian Electoral Commission, the Australian Federal Police, and intelligence agencies, to advise the Australian Electoral Commissioner on any issues that could compromise electoral integrity. Its scope covers everything from cyber- and physical security to disinformation campaigns and interference. While the EIAT's mandate does not include direct election management, it plays a crucial advisory role in ensuring the integrity of elections. One key lesson from the EIAT is its collaborative, cross-agency approach, which provides a consolidated view of electoral threats.

Australia's **Defending Democracy Unit**, launched in 2022, operates as a permanent unit within the Australian Electoral Commission. Its primary objective is to counter threats to electoral integrity, including disinformation and electoral interference. The unit's proactive education campaigns around electoral events are aimed at strengthening public awareness and resilience against false information, demonstrating the importance of public outreach in modern electoral systems.

In Canada, the **Plan to Protect Democracy** was introduced prior to the 2019 federal elections to strengthen its electoral system against external and internal threats[20]. The cornerstone of this effort is the **Security and Intelligence Threats to Elections (SITE)** Task Force, which coordinates intelligence from key security bodies in Canada, such as the Canadian Security Intelligence Service (CSIS), Global Affairs Canada, and the Royal Canadian Mounted Police. SITE focuses on monitoring and countering foreign interference, disinformation, and cyber threats. This approach shows the value of multi-agency coordination in responding to increasingly complex electoral challenges.

Estonia's State Electoral Office stands out for its innovative approach in addressing the influence of disinformation. Estonia, leveraging its advanced digital infrastructure, established an inter-agency taskforce in 2016 to counter the spread of false messaging aimed at disrupting its elections. What distinguishes Estonia is its "network" model, which involves light resources but focuses on collaboration between government agencies, civil society, social media platforms, and traditional media. Estonia's success also highlights the value of civics education and media literacy, integrated into its strategy for strengthening public trust.

CASE STUDY: VIGINUM - FRANCE'S VANGUARD AGAINST FOREIGN DIGITAL INTERFERENCE

The Vigilance and Protection Service Against Foreign Digital Interference (VIGINUM), established in 2021, is France's specialised agency for countering foreign manipulation of public discourse, particularly during elections. Operating under the General Secretariat for Defence and National Security (SGDSN), VIGINUM is tasked with detecting and disrupting foreign interference while avoiding domestic censorship, ensuring the protection of democratic freedoms[21].

The agency's success is driven by its interdisciplinary team, involving analysts, data scientists, and digital media specialists. As of 2022, VIGINUM operated with an annual budget of €12 million and aimed to staff 65 employees[22], including analysts, data engineers, and media experts.

VIGINUM's structure is designed for focus and agility. Its interdisciplinary team of analysts, data scientists, and digital media experts leverages open-source intelligence (OSINT) from media, social platforms, and other public data to detect disinformation patterns in real time. Advanced technical tools enable the agency to monitor and analyse digital content, identifying threats and attributing them to foreign actors[23]. To maintain trust, VIGINUM operates under the oversight of the French National Commission for Information Technology and Civil Liberties (CNIL) and an independent ethical and scientific committee, ensuring compliance with privacy laws and democratic values.

The Matryoshka Campaign, revealed in VIGINUM's 2024 report[24], illustrates the sophistication of modern disinformation tactics as well as VIGINUM's operating model. As VIGINUM reports, this Russian-led operation deployed "seeders" to introduce false narratives and "quoters" to amplify them, creating the illusion of organic discourse. VIGINUM's ability to expose and disrupt the campaign highlights its effectiveness. By analysing digital patterns and tracing disinformation flows, the agency attributed the operation to foreign actors and neutralised its impact.

VIGINUM's targeted approach — focusing on electoral periods and foreign interference — ensures efficient resource allocation while avoiding mission creep. Its integration within the SGDSN allows for seamless coordination with France's national security apparatus, enhancing its capacity to respond to threats swiftly. By balancing technical expertise with ethical oversight, VIGINUM offers a robust model for safeguarding democratic processes against foreign digital manipulation.

VIGINUM provides an encouraging blueprint for democracies to counter digital interference effectively. Its focus on in-house expertise, transparency, adaptability, and alignment with democratic principles demonstrates how to defend elections and public discourse in an era of increasingly sophisticated disinformation campaigns.

THE LIMITS OF EXISTING INSTITUTIONS AND THE CHALLENGE OF SPEED

Electoral integrity institutions across various democracies have made strides in addressing traditional threats like foreign interference and disinformation. However, in today's rapidly evolving digital landscape, even these institutions may not be fully prepared to tackle the challenges posed by synthetic content generated by advanced AI. The rise of deepfake technology, AI-generated imagery, and automated disinformation campaigns fundamentally shifts the nature of the threat, with speed now being the defining factor.

The speed and scale of synthetic content production pose a serious challenge. Unlike traditional disinformation, which required manual crafting and dissemination, synthetic content can be generated and distributed algorithmically in minutes. This enables tactics like "flooding the zone," where vast amounts of disinformation overwhelm platforms, shaping public opinion before any intervention can occur. Many of these institutions are not equipped to handle the sheer volume and velocity of these AI-enabled disinformation attacks.

One issue is that most current institutions remain focused on traditional disinformation and foreign interference — a purpose for which they were designed — without fully addressing the complexities of AI-driven manipulation. Synthetic content introduces new vectors of attack that require institutions to adapt not just their monitoring capabilities, but also their operational frameworks.

Estonia's State Electoral Office is an example of a forward-thinking, lightly resourced institution that has successfully used a "network" approach to combat traditional disinformation. However, even this model will likely need enhancement in an era of synthetic content, where AI-generated material demands faster, more sophisticated detection methods.

The key is not just having technology, but integrating it effectively with human judgement. Synthetic content like deepfakes often requires contextual understanding that AI alone cannot provide. However, relying too heavily on human analysis can slow response times, allowing disinformation to spread unchecked. The right balance between automated detection and human oversight is essential, with AI flagging risks and human analysts providing the nuanced contextual judgement needed for triaging and intervention.

Many existing institutions struggle with this balance. While platforms like Meta pioneered AI to detect *coordinated inauthentic behaviour.*However, whether Meta will continue using this method alongside its newly announced community notes-based approach remains to be seen. Similarly, electoral integrity bodies need to similarly incorporate advanced technologies to manage AI-generated disinformation. Without these tools, institutions risk being overwhelmed by the sheer volume of synthetic content and left playing catch-up.

LESSONS FROM OTHER FIELDS: MODERN INSTITUTIONS FOR MODERN PROBLEMS

'Modern Institution' (Generated by Dall-E)

History shows that pre-existing institutions often struggle to address new technological challenges. This is why new ones have so often had to be created — whether for human fertilisation or AI Safety, space exploration or clean energy.

Existing institutions can adapt but often they become locked into routines which then persist because of entrenched interests, sunk costs, and resistance to change[25].

Traditional institutions responsible for electoral integrity, though crucial, may not be equipped to handle the sophisticated nature of AI-driven misinformation campaigns. This is why we argue for new or updated institutions with capabilities better suited to handling current and likely technological threats.

In the next few sections we give some examples that show how this can be done, and how often new institutions are needed when there are not only new tasks, but also a new ethos needed.

NEW INSTITUTIONS FOR NEW TASKS (ICANN)

The creation of the Internet Corporation for Assigned Names and Numbers (ICANN) in 1998 is a prime example of how innovative institutions can successfully address emerging technological issues. This institution was a pioneering solution to the chaos that could have ensued as the internet expanded exponentially. ICANN was not merely a modification of existing internet governance structures but a novel institution, meticulously designed to manage the global domain name system and ensure the internet's stability and security. One of ICANN's key innovations is its use of "rough consensus" as a decision-making process. This method involves open debate among experts to refine solutions until they have the broad support of participants. This approach allows ICANN to remain adaptive and effective, leveraging the collective expertise of a diverse, global community to make credible and legitimate decisions.

ICANN's adaptability was clearly demonstrated during the 2022 conflict between Russia and Ukraine, when Ukraine requested the disconnection of Russia from the internet to limit propaganda. ICANN, along with the Regional Internet Registries (RIRs), declined to intervene, reaffirming their commitment to neutrality and non-interference. This response underscores ICANN's foundational norms of equal treatment and operational focus, ensuring that the internet remains a tool for communication free from political influence (Sowell, forthcoming).

ICANN's success also lies in its administrative authority, which ensures the integrity of the rough consensus process. This authority maintains unbiased and operationally sound decision-making, grounded in practical experience rather than external political pressures. The RIPE NCC, one of the RIRs, exemplifies this by guaranteeing equal treatment for all internet service providers, reinforcing the credibility of the rules established through rough consensus.

ONE-SIZE RARELY FITS ALL (FSB)

Many institutions struggle when faced with a new, and complex challenge.

A good example was the 2008 global financial crisis which revealed the inadequacies of existing regulators. Many financial regulatory bodies were built on frameworks designed for simpler, less interconnected markets, which rendered them inadequate for the complexities of modern global finance. For instance, the U.S. Securities and Exchange Commission (SEC) and other national regulators focused primarily on domestic markets and traditional banking institutions, overlooking the burgeoning shadow banking system and the complex, interlinked financial products like mortgage-backed securities and credit default swaps. The crisis revealed significant gaps in oversight and coordination among national regulators, leading to widespread economic instability.

Unlike its predecessors, the Financial Stability Board (FSB) was designed specifically to handle the intricacies of the global financial system. Although it has faced criticism for some aspects of its design, such as its lack of enforcement power and limited transparency, it has introduced some key innovative methods. These include *stress testing* financial institutions to assess their resilience to economic shocks or enhanced international cooperation by bringing together national financial authorities and standard-setting bodies. This tailored approach allowed the FSB to more effectively address the interconnected and dynamic nature of global finance, stabilising the system in ways that a path dependent, or off-the-shelf regulatory framework could not.

NOT ALL NEEDS TO BE NEW (FDA)

It's not always necessary to create something new, and even new institutions can learn from older ones.

For example, the Ada Lovelace Institute argues that GenAI regulation can learn from the US Food and Drug Administration (FDA)[26]. The FDA's core principles such as continuous engagement with developers, wide-ranging information access, and placing the burden of proof on developers to demonstrate safety and efficacy ensure that products undergo thorough scrutiny before reaching the market and are continuously monitored post-market to address emerging risks. Applying a similar framework to GenAI could help mitigate potential harms, ensuring these technologies are developed and deployed in a manner that prioritises public safety and trust.

Drawing from these examples, addressing the threat of synthetic content to electoral integrity is likely to require institutions with a similarly creative and tailored approach. With the context set, let us now turn our attention to the Framework itself.

THE FRAMEWORK: ELECTORAL INTEGRITY INSTITUTIONS IN A WORLD OF AI AND SYNTHETIC CONTENT

The following sections are a guide to the key design challenges likely to be encountered when creating institutions focused on electoral integrity. Each section offers a range of options for institutional architects to consider, complemented by relevant case studies and lessons learned from existing models.

The remainder of this paper delves into the core steps of the logic model underpinning our electoral integrity institution. This aims to construct a blueprint of an institution designed to effectively confront the challenges of synthetic content. See Figure 5 below, which encapsulates the high-level institutional architecture.

SET THE RIGHT FOUNDATIONS

Setting up the essential attributes, systems, and processes right is crucial for the institution's success. This begins with the fundamentals — laying the groundwork for everything else to build upon.

THE FUNDAMENTALS: TASK, PEOPLE, ETHOS

Task

Clearly defining the task or purpose is the foundation of any organisation. The task dictates the organisation's structure, capabilities, and scale. Without clarity on the task, resources can be misallocated, and efforts may lack direction. Knowing the purpose and all the key activities that contribute towards it, ensures all efforts are aligned and focused, much like how a laser, when focused, can cut through steel. Below are some of the key considerations in relation to defining the task of this institution:

— *Establishing Boundaries and Scope:*

The remit of the institution must be clearly defined to specify where its responsibilities begin and end. This includes deciding whether the focus will be narrow—such as safeguarding the electoral process during key periods like the lead-up to an election, election day, and its immediate aftermath—or broader, addressing long-term harms to societal discourse and public deliberation year-round.

A focused scope may target the "tip of the iceberg," dealing with acute threats to electoral integrity. Conversely, a broader remit might include the "body of the iceberg," tackling incremental yet pervasive challenges to epistemic security. Striking this balance is critical, as it shapes the institution's operational design and ensures efforts are complementary to existing institutions without duplication.

— *Synthetic Content as a Strategic Driver*:

Among the spectrum of disinformation threats, synthetic content—such as deepfakes and AI-generated misinformation—stands out as a distinct and urgent challenge. Synthetic content encapsulates many of the risks inherent in broader disinformation, with its scale, speed, and sophistication presenting unique vulnerabilities. Focusing on synthetic content allows for the development of specialised tools and interventions, providing a practical entry point to address a rapidly growing threat.

— *Elections at the Core:*

Electoral processes lie at the heart of democratic systems and represent a natural focal point for this institution. Elections are not only critical to public trust and governance, but are also where the impacts of synthetic content are most visible and measurable. By prioritising electoral integrity, the institution can target a domain where synthetic content has the potential to inflict significant harm while providing a foundation for broader applications in the future.

Nevertheless, for the purpose of this paper, we propose a rough sketch of the institution and its key tasks and functions. Broadly, we believe the institution will need to:

1. **FACILITATE collaboration among stakeholders:** Bringing together diverse stakeholders and fostering collaboration towards shared goals.

2. **SCAN the digital space:** Mobilising key human and computational resources and tapping into the collective intelligence of the people and machines to identify adversarial content.

3. **ASSESS effectively and impartially:** Balancing speed with responsibility to impartially triage content.

4. **ACT with power and accountability:** Executing efficiently with strong mandate and democratic checks and balances.

5. **LEARN via feedback loops:** Implementing feedback loops to rapidly learn and incorporate new information.

People

The success of an organisation hinges on its people. This involves recruiting individuals with the right expertise, fostering a diverse mix of skills and backgrounds, and ensuring a balance of complementary abilities within teams. For the present institution, this may mean bringing together experts in misinformation, cybersecurity, media, legal compliance, and public policy. This involves not just finding capable individuals but ensuring a diverse mix of backgrounds and experiences. Effective teams balance complementary skills and mindsets, blending strategic vision with attention to detail. In the following sections, this paper delves into further detail regarding the types of expertise that are required (e.g., Foresight; OSINT) and how to embed them within the institution.

Ethos

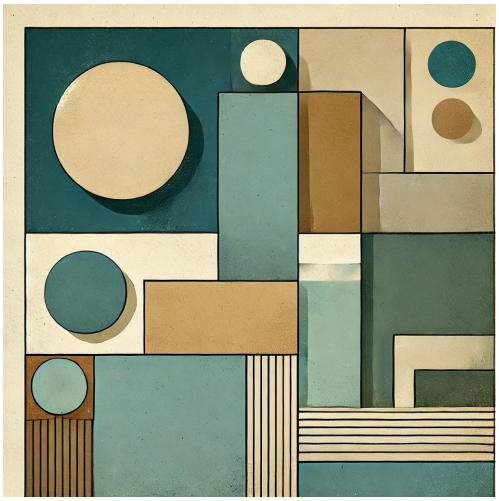
Finally, an organisation's ethos, or core values, underpin its identity and actions. These values guide decision-making and behaviour, shaping the organisation's culture. A strong ethos is not just a mission statement but a guiding principle that shapes what the organisation does and does not do. Effective organisations ensure their ethos is lived daily, reinforced through actions and decisions that reflect their core values.

Similarly to the Task, the right ethos of the institution will depend on the local context. However, for the purposes of this paper, we sketch out key values that we can reasonably expect to apply. These values are: *transparency*, *collaboration*, *accountability*, *promptness*, and *impartiality*.

The ethos should be deeply embedded in daily operations, ensuring that every action and decision reflects these core principles. This consistency helps the institution maintain integrity and effectiveness, as well as public trust and credibility.

History and experience have taught us that setting the institution's basics, particularly Task/Purpose and Ethos through *co-creation*, involving relevant stakeholders in its development, is far more effective than top-down *dictation*. This inclusive approach not only ensures that the values represent the collective will of the people involved with the organisation, delivering on its mission day in day out; it also fosters a sense of ownership, ensuring that people identify with, buy into, and embody these values, particularly during challenging times.

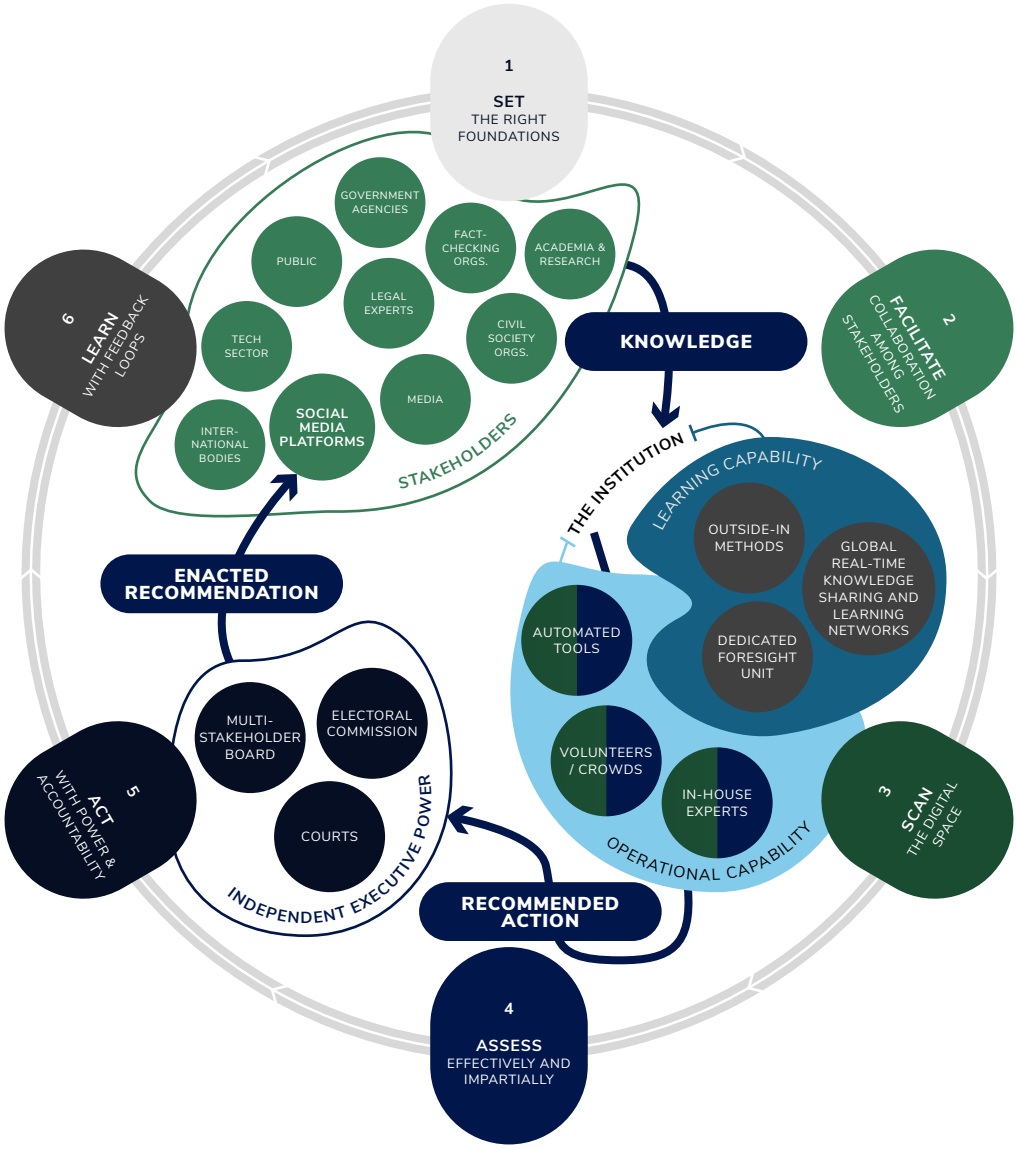
THE DESIGN MUST MIRROR THE FUNDAMENTALS



'Design Mirroring Fundamentals in Concrete Art Style' (Generated by Dall-E)

To effectively address the complex challenge of protecting electoral integrity against the harms of synthetic media, the organisational design must embody the multifaceted ethos of the institution. Given its role of a hub, these are likely to consist of diverse values such as transparency, collaboration, accountability, promptness, and impartiality. These values can often be in a dialectical conflict (e.g., collaboration and the need to involve people in decisions can clash with the need for promptness). To effectively incorporate and manage this complexity, the design needs to integrate elements from various organisational approaches — hierarchical, egalitarian, and individualistic — to create a robust and effective institution. The design should be ideology-agnostic, focusing on the institution's core purpose. In this case, it must reflect its hub-like, associative role while balancing speed, accountability, and impartiality in decision-making and execution.

Figure 1. How the Electoral Integrity Institution operates: a high-level visualisation



FACILITATE COLLABORATION AMONG STAKEHOLDERS

To protect electoral integrity across diverse contexts, the institution must function as a hub that brings together a wide array of stakeholders effectively. Achieving this requires designing intelligent mechanisms and decision-making processes that facilitate meaningful interactions. In this section, we present a menu of actionable options for designing the institution, with a particular focus on aligning the often-conflicting incentives of its diverse stakeholders.

'Harnessing Tensions' (Generated by Dall-E)

Inspired by emergency response hubs, which coordinate diverse actors during crises, we propose a similar model for safeguarding electoral integrity. Emergency hubs bring together first responders, medical teams, government agencies, non-profits, and community volunteers under a unified framework to ensure cohesive and efficient responses. Without such coordination, fragmented efforts can lead to inefficiencies and critical delays. This principle translates directly to the context of elections: in a world where synthetic content and misinformation pose growing threats, a well-designed hub ensures alignment, avoids duplication, and enables swift, effective interventions.

Modern electoral threats are wicked and complex, requiring **multistakeholder collaboration** to address effectively. While this approach has gained prominence, it also faces significant challenges, such as balancing competing interests, overcoming trust deficits, and aligning incentives. Addressing these challenges requires a deep understanding of stakeholder dynamics and careful incorporation of lessons from successful initiatives.

This section outlines key principles, best practices, and innovative options for designing an institution capable of facilitating effective collaboration among a diverse pool of stakeholders.

WHO ARE THE STAKEHOLDERS?

An institution dedicated to protecting electoral integrity must unite a diverse array of stakeholders (see *Appendix A and B for practical tips on how to design a collaborative environment*). Each group plays a vital role in ensuring comprehensive and effective action. The following is an example, as any specific pools of stakeholders will depend on the local context.

— *Government Agencies:*

Government agencies are fundamental. Election commissions oversee the electoral process and implement policies to maintain its integrity. Cybersecurity agencies protect against cyber threats and misinformation campaigns targeting elections. Law enforcement investigates and addresses illegal activities related to electoral fraud and misinformation.

— *Technology Companies:*

Technology companies are vital allies. Social media platforms like Meta, X, and Google are on the frontline of monitoring and controlling the spread of misinformation–though Meta's move to follow X in shifting away from human fact-checkers in favour of community notes signals a broader trend of retreating from dedicated fact-checking efforts. Technology providers develop and supply tools for fact-checking, data analysis, and cybersecurity, equipping the hub with necessary resources to identify and counteract false information effectively.

— *Fact-Checking Organisations:*

Independent fact-checkers and collaborative networks are essential. Organisations like Full Fact, PolitiFact, or Logically verify information and debunk false claims. Collaborative initiatives, such as the International Fact-Checking Network (IFCN), facilitate cooperation and resource sharing among fact-checking entities, enhancing their collective capacity to combat misinformation. Tapping into the capabilities of these organisations may be essential to success.

— *Media Outlets:*

Media outlets are indispensable for disseminating accurate information. Both traditional and digital news organisations report on electoral processes and correct misinformation. Journalist associations provide training and resources, equipping journalists with the skills needed to identify and report false information accurately.

— *Academic Institutions and Researchers:*

Academic institutions and researchers offer deep insights. Universities and research institutes, such as think tanks, conduct studies on misinformation, electoral integrity, and related technologies. Data scientists and social scientists analyse data trends and the impact of misinformation on public opinion, providing valuable insights that inform strategies and actions.

— *Civil Society Organisations:*

Civil society organisations, including NGOs and community groups, mobilise grassroots efforts to monitor and report on electoral issues. They engage in advocacy, voter education, and monitoring of electoral processes, ensuring community involvement and vigilance.

— *International Bodies:*

International bodies such as OECD, or election observation missions from organisations like the OSCE or The Carter Center observe and report on the integrity of elections globally. These bodies provide an external perspective and validate the fairness and transparency of electoral processes.

— *Legal Experts:*

Legal experts, including election law specialists and policy advisors, offer guidance on election laws and regulations. Their expertise is crucial for interpreting legal frameworks and ensuring that actions taken to protect electoral integrity comply with the law.

— *The Public:*

Engaging the public is crucial. Educating voters about misinformation and equipping them with tools to identify false information is essential. Volunteers play a significant role in monitoring, reporting, and verifying information, contributing to a community-driven approach to safeguarding electoral integrity.

MAKING MULTISTAKEHOLDER GOVERNANCE WORK

Multistakeholder governance has emerged as a preferred model for addressing global challenges that require collaboration across governments, corporations, civil society, and other actors. However, these models are not without risks: corporate capture, misaligned incentives, and inefficiencies can undermine their effectiveness[27].

However, there have been promising examples of multistakeholder collaboration, such as Adobe's Content Authenticity Initiative (CAI). CAI brings together organisations like Amazon, Meta, news outlets, and NGOs under the Coalition for Content Provenance and Authenticity (C2PA) to develop open standards for content verification. This effort has proven vital in enhancing transparency. Recently, Adobe launched a free web app to help creators protect their work and verify content authenticity, further democratising access to content authentication tools. Initiatives like CAI set a strong precedent for fostering collaboration and innovation in addressing synthetic content.

By examining key principles, leaning on lessons from existing initiatives and insights from key academic texts, this section provides suggestions for designing robust a multistakeholder governance model.

— *The Importance of Clear Mandates and Goals:* One of the most critical factors for the success of multistakeholder governance is the establishment of a clear mandate. Without a focused purpose, governance structures risk inefficiency and diluted impact. As experts on multistakeholder governance have warned, vague or overly broad mandates allow dominant actors to exploit ambiguities to serve their interests[28].

The **Christchurch Call** offers a compelling example of how clear goals can drive meaningful progress. Established in 2019 to combat online extremism, the Call sets out specific, actionable objectives, such as improving algorithmic transparency and developing rapid content takedown systems[29]. These goals are publicly reported, fostering both transparency and accountability. Electoral integrity institutions can emulate this approach by defining objectives such as deploying detection capabilities mentioned in this paper or working with social media platforms to reduce the prevalence of electoral disinformation by measurable percentages.

— *Balancing Representation to Avoid Dominance:* Effective multistakeholder governance relies on achieving balanced representation while addressing power dynamics that can undermine its legitimacy. Drawing on the analysis of a leading expert on transnational governance[30], it becomes clear that multistakeholder structures like the **Forest Stewardship Council** (FSC), while often praised for their inclusiveness, face challenges in preventing disproportionate influence from powerful stakeholders. The FSC's tripartite structure, which allocates equal voting power to economic, environmental, and social chambers, is designed to ensure no single group dominates decision-making. However, in practice, economic actors frequently exert outsized influence, exposing a gap between structural ideals and operational realities.

This critique extends to how inclusiveness is used as a legitimacy strategy. The FSC has, at times, relied on inclusiveness more as a symbolic gesture — what has been termed "window dressing" — rather than fostering genuine equity among stakeholders. Such practices risk reducing representation to superficiality, undermining the intended balance of power. Furthermore, once governance structures like the FSC are established, they often become resistant to change due to path dependency, which limits their ability to adapt to evolving challenges or respond to criticisms.

These lessons offer critical guidance for electoral integrity institutions. While quotas for corporate, civil society, and academic representation are essential, they must be paired with mechanisms that actively address power imbalances. Rotating leadership positions, embedding transparent decision-making processes, and instituting independent oversight are all crucial steps to ensure that representation remains meaningful and effective[31]. Moreover, adaptability must be built into the institution's design, enabling periodic review and refinement of structures to address inequities and evolving challenges, rather than relying on initial frameworks to uphold legitimacy.

— *Securing Financial Independence:* Funding sources are a key vulnerability for multistakeholder governance models. While government funding is often the foundation, additional contributions from private actors must be carefully managed to avoid undue influence. The **Global Fund** to Fight AIDS, Tuberculosis, and Malaria demonstrates best practices in this regard. Its pooled funding model caps individual contributions and employs independent auditors to oversee financial flows, maintaining accountability while preventing donor-driven agendas (Global Fund, 2022).

For electoral integrity institutions, diversifying funding sources while safeguarding decision-making independence is critical. Additional contributions from private entities should be managed transparently, with mechanisms like capped contributions or oversight by independent financial bodies being important options to consider.

CASE STUDY: HOW TO ALIGN DIVERSE STAKEHOLDERS AROUND A COMMON GOAL? LESSONS FROM COP

Given the diversity of stakeholders, many of which may be driven by conflicting incentives, the success of an institution safeguarding election integrity hinges on effectively aligning these stakeholders around the shared goal of safeguarding electoral integrity.

One exemplary institution that has navigated this challenge since its inception is the United Nations Climate Change Conference (COP). COP serves as a high-profile model for uniting diverse stakeholders with conflicting incentives around a common objective. While COP's impact has had its challenges, the strategies it has employed can still provide valuable insights for any institution aiming to address a complex challenge in a multistakeholder, multi-incentive environment. Here are some key methods and tools used by COP:

Structured Negotiation Sessions: COP conferences involve intense negotiation sessions, where government representatives draft and agree on climate action commitments. For instance, during COP21 in 2015, these negotiations led to the Paris Agreement, with nearly 200 countries committing to limit global warming. This process exemplifies how structured, formal discussions can balance different priorities and capabilities, ensuring fair and actionable agreements for all parties. However, COP teaches us that these sessions can be protracted and sometimes result in compromises that dilute the impact of the agreements. The lesson here is to strive for timely negotiations while maintaining the integrity of commitments.

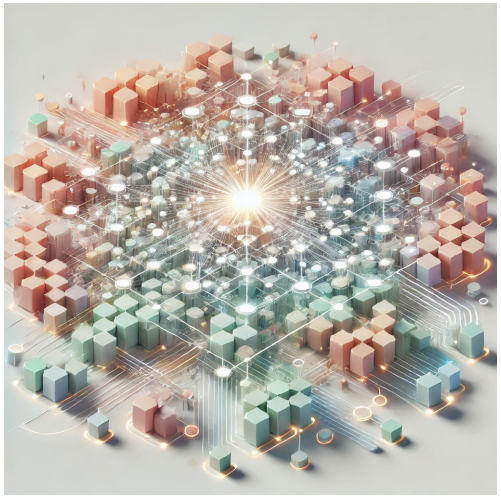
Collaborative Side Events: Side events at COP conferences include workshops, panel discussions, and presentations hosted by various organisations. These events foster knowledge sharing and innovation. For example, COP26 featured discussions on renewable energy, sustainable agriculture, and climate finance, allowing stakeholders to exchange ideas and forge new partnerships. This demonstrates the importance of creating spaces for informal collaboration alongside formal negotiations. However, COP teaches us that the plethora of side events can lead to information overload and dilute the focus on core negotiations. Institutions should carefully curate side events to ensure they complement rather than overwhelm the primary agenda.

Focused Working Groups: COP utilises working groups composed of experts from diverse fields to tackle specific issues such as carbon markets and technology transfer. These groups delve deeply into technical details and propose solutions that feed into the larger negotiation process. This approach highlights the effectiveness of task-specific teams in addressing complex problems with targeted expertise. The lesson from COP is that while effective, these groups can sometimes struggle with coordination and integrating their findings into the broader COP agenda. Effective coordination and clear integration mechanisms are essential to maximise the impact of specialised working groups.

Transparent Plenary Sessions: Plenary sessions at COP ensure that all participants stay informed about negotiation progress and can contribute to high-level decision-making. Updates, progress reports, and major decisions are shared transparently, building trust among stakeholders. This underscores the value of maintaining transparency and open communication in fostering cooperation and collective action. However, COP also teaches us that the sheer scale of participation can make these sessions unwieldy and occasionally slow down decision-making processes. Ensuring streamlined, efficient plenary sessions without sacrificing transparency is a crucial lesson.

By integrating these tools, COP conferences unite diverse stakeholders around the shared goal of combating climate change. These strategies balance differing interests, foster a spirit of cooperation, and promote collective action, offering a model to learn from for any institution seeking to address complex, multistakeholder challenges.

WHO OWNS THIS? COORDINATING ACROSS FRAGMENTED INSTITUTIONS:



'Coordination Across Fragmented Institutions' (Generated by Dall-E)

Disinformation related to elections occupies a fragmented and often ambiguous regulatory space, intersecting with multiple institutions. Electoral commissions oversee the voting process but may lack the resources or expertise to tackle digital threats. Media regulators address harmful content but often focus on traditional broadcasters, leaving online platforms under-regulated. Cybersecurity agencies defend against technical breaches but may not prioritise the narrative and psychological dimensions of disinformation. Intelligence services monitor foreign influence but operate with secrecy that can hinder broader collaborative efforts. These overlapping mandates create a governance gap, where responsibility is diluted, and accountability is unclear.

A stark illustration of this challenge comes from the 2020 UK Russia report[32], which

highlighted how institutions, faced with Kremlin interference in the Scottish referendum and Brexit campaigns, failed to act decisively. Instead, agencies pointed fingers at one another, exposing the systemic failure of a fragmented approach. Such governance gaps leave electoral processes vulnerable to manipulation and erode public trust.

This framework proposes an **Electoral Integrity Institution** designed to fill this gap by driving collaboration and taking ownership of the problem. Rather than duplicating the efforts of existing bodies, this institution acts as the central coordinating hub, ensuring alignment and accountability across relevant stakeholders. Here's how this collaboration could be structured:

1. *Centralised Coordination Through Defined Mandates:*

   The institution's first task is to establish clear lines of responsibility. By mapping the roles and capacities of existing agencies—such as electoral commissions, media regulators, cybersecurity agencies, and intelligence services—it can delineate where its own remit begins and ends. This clarity prevents duplication of efforts and reduces the risk of gaps in coverage. For example, while intelligence agencies might detect foreign interference campaigns, this institution would focus on translating such intelligence into actionable strategies for countering disinformation in public narratives.

2. *Driving Collaboration Through Formalised Agreements:*

   To overcome institutional silos, the institution can implement memorandums of understanding (MOUs) with each stakeholder. These agreements formalise collaboration, ensuring that each agency contributes its unique expertise while maintaining accountability for their respective responsibilities. For instance, the institution could work with cybersecurity agencies to monitor online manipulation infrastructure while leveraging media regulators to address platform compliance with disinformation policies.

3. *Owning the Problem Through Operational Leadership:*

   What distinguishes this institution is its operational leadership. By establishing itself as the authoritative body for election-related disinformation, it consolidates fragmented efforts into a coherent strategy. This leadership role involves not just coordination but proactive problem-solving — identifying emerging threats, mobilising resources, and issuing timely directives. For instance, in the face of a coordinated disinformation campaign, the institution would lead cross-agency rapid response efforts, ensuring consistency and effectiveness across sectors.

4. *Facilitating Information Sharing and Transparency:*

   Fragmentation often stems from poor communication and information silos. This institution would establish a centralised data-sharing platform, enabling secure and timely exchange of intelligence between agencies. Drawing lessons from successful inter-agency collaborations, such as the Joint Terrorism Task Force in the United States, the platform would ensure that relevant information is accessible to all stakeholders without unnecessary bureaucratic delays.

5. *Accountability and Public Trust:*

   To counter perceptions of turf wars or passivity, the institution must operate with transparency and accountability. Regular public reporting, outlining its coordination efforts and outcomes, reinforces its role as the trusted guardian of electoral integrity. By actively engaging with civil society and the media, the institution builds public confidence in its ability to own and address the disinformation challenge.

By positioning itself as the central authority for electoral disinformation, this institution addresses the governance gap that has long been plagued by fragmented efforts. This approach transforms the disinformation problem from an issue of diffused responsibility into one of collective, coordinated action.

NAVIGATING CROSS-BORDER CHALLENGES: MULTILATERAL GOVERNANCE WITH NATIONAL FOCUS

Synthetic content and its potential impact on electoral integrity are inherently global issues. Digital platforms operate across borders, and disinformation campaigns frequently exploit jurisdictional gaps to maximise their impact. Such cross-border challenges demand a coordinated international response. However, the proposed Electoral Integrity Institution remains **primarily focused on national applicability**, addressing the unique needs of domestic electoral systems. To ensure effectiveness in this globalised context, the institution must strategically engage with multilateral governance frameworks while maintaining its core national focus.

The borderless nature of synthetic content allows disinformation campaigns to thrive by targeting weak points in international coordination. Malicious actors often exploit countries with less stringent regulatory environments or focus on diaspora communities — populations living outside their country of origin but maintaining strong ties to it. Diasporas can serve as critical conduits for disinformation, influencing political discourse in both host and home countries. This dynamic underscores the need for a robust multilateral approach that integrates national initiatives into a cohesive global strategy.

Key Multilateral Efforts for Collaboration:

To address these challenges, the institution can align with and complement existing multilateral frameworks. Several key organisations provide established mechanisms and expertise:

— **The OECD's Digital Governance Initiatives:**

  The OECD's work in AI ethics and digital governance is foundational for fostering international cooperation on synthetic content. Its frameworks, such as the OECD AI Principles, provide standards for responsible technology use and cross-border policy coordination. By engaging with the OECD, the institution can adopt and customise these standards to address the unique vulnerabilities of electoral systems while contributing to the global effort against disinformation.

— **NATO's Hybrid Threat Response:**

NATO's extensive experience in countering hybrid threats — including state-sponsored disinformation — makes it an essential partner. The proposed institution could collaborate with NATO to share intelligence, develop synthetic content detection capabilities, and run joint simulations to test electoral resilience against cross-border interference. NATO's security-first approach complements the institution's electoral focus, ensuring alignment without redundancy.

— **UNESCO's Media and Information Literacy Programmes:**

UNESCO's global leadership in fostering media literacy provides a critical foundation for societal resilience against misinformation. Partnering with UNESCO would enable the institution to integrate synthetic content detection and awareness into existing media literacy programmes, focusing on vulnerable groups like diaspora communities. Such collaboration would ensure consistent, high-impact messaging across jurisdictions.

— **The European Digital Media Observatory (EDMO):**

EDMO specialises in disinformation research, fact-checking, and promoting media literacy within the European Union. Its partnerships with academia, civil society, and fact-checking networks make it a valuable ally in developing cross-border strategies for detecting and combating synthetic content.

— **The G7 Rapid Response Mechanism (RRM):**

The G7 RRM was established to counter foreign interference in democracies. Its focus on information-sharing and coordinated responses makes it an ideal partner for addressing synthetic content campaigns that target multiple countries simultaneously. Collaboration could involve joint research initiatives and the alignment of rapid response protocols.

Despite the importance of multilateral engagement, the institution's primary focus remains on national applicability. Electoral systems are deeply embedded in specific political, cultural, and legal contexts, requiring tailored approaches. To balance these priorities effectively, the institution should:

— *Adapt Global Standards Locally:* Integrate international standards for while ensuring alignment with domestic regulatory frameworks.

— *Strengthen Cross-Border Coordination Without Ceding Sovereignty*: Establish liaison mechanisms to collaborate with multilateral bodies while retaining decision-making authority at the national level.

— *Address Diaspora-Specific Vulnerabilities:* Develop targeted strategies to combat synthetic content aimed at diaspora communities, leveraging their influence in both host and home nations. This could include public awareness campaigns and monitoring tools tailored to diaspora communication channels.

The institution could therefore be adopt a dual-focused governance model, serving as both a national authority and a multilateral participant. A dedicated unit would be tasked with liaising with organisations like the OECD, NATO, UNESCO, EDMO, and the G7 RRM, ensuring collaborative efforts enhance national initiatives rather than dilute them. This approach allows the institution to address cross-border challenges effectively while safeguarding its primary commitment to domestic electoral integrity.

SCAN THE DIGITAL SPACE

The sheer volume of online content is already staggering, but as many experts predict[33], the arrival of genAI and synthetic content mean it will continue to increase exponentially. To effectively monitor this deluge of information, any institution that aims to sense the landscape needs to bolster their ranks with a large number of volunteers, be equipped with cutting edge technological tools, and be able to make the most out of the collective intelligence these resources offer.

In this section, we explore ways of recruiting and mobilising such resources, how to harness the collective intelligence of the humans and machines, why synthetic threats need to be treated as a dissemination problem, not a content problem, and what lessons can be learned from collaborative fact-checking initiatives.

BALANCING RESOURCE DEMANDS WITH EXPERTS AND VOLUNTEERS

There is no hiding it. The tasks of the electoral integrity institution are exceptionally resource-intensive, requiring both technical sophistication and operational scale. Traditional electoral commissions are often under-resourced for such challenges, lacking the expertise and infrastructure to manage the sheer complexity and volume of disinformation threats. This framework addresses these limitations by proposing a model that integrates in-house expertise with strategically deployed volunteer networks.

In-house expertise is indispensable for tasks requiring high-level technical and contextual knowledge, such as interpreting threat intelligence, and coordinating multi-agency responses. These roles demand specialists with deep expertise in areas such as generative AI, information warfare, and electoral processes, making investment in skilled personnel a non-negotiable requirement. Without this core, an institution risks being reactive and prone to missteps in identifying and addressing threats.

To augment this professional backbone, volunteers can provide valuable support in key areas. Drawing lessons from successful models like open-source intelligence (OSINT) communities, volunteers enable institutions to expand their capacity without incurring unsustainable costs. For instance, volunteer contributions can be particularly effective in rapidly assessing large datasets, monitoring niche online spaces, or flagging emergent narratives for further analysis.

A model that combines expert-led operations with a well-structured volunteer network not only optimises resource allocation but also addresses a critical gap in traditional electoral institutions: the ability to respond with speed and scale to disinformation campaigns.

MOBILISING VOLUNTEERS

Volunteers can be crucial in combating misinformation, particularly during elections. This is because they provide a diverse, decentralised, and scalable workforce capable of assessing content quickly and efficiently. Studies have shown that crowd workers can be remarkably effective at identifying false information, often achieving levels of accuracy comparable to experts[34]. Their varied backgrounds and viewpoints mean they can collectively evaluate content more holistically, catching nuances and contextual details that automated systems or single experts might miss. Moreover, volunteers are often motivated by a genuine desire to contribute to the integrity of public discourse, which enhances the quality and reliability of their assessments. This collective intelligence approach, combining the efforts of many individuals, has proven to be an incredibly powerful tool in rapidly flagging and correcting misinformation, making crowd-sourced fact-checking an essential part of any comprehensive strategy to counteract falsehoods online.

Open source evidence has proven particularly effective in forensically analysing footage from conflict zones, such as Gaza, to verify authenticity and expose misleading content[35]. Volunteers and crowd workers, often organised through initiatives like the TRUE project, meticulously examine details like shadows, landmarks, and metadata to confirm whether videos are genuine or manipulated. This grassroots approach has been instrumental in debunking false narratives and ensuring that the truth from war-torn regions reaches the public, further demonstrating the power of collective intelligence in fact checking even the most complex and contentious situations.

Key Considerations: When recruiting volunteers, time and again the collective intelligence literature has shown that it is essential to ensure the group is as diverse as possible, reflecting different backgrounds, perspectives, and skills. Diversity strengthens the institution's ability to detect and understand a wide range of misinformation tactics and impacts. Additionally, volunteers should be equipped with key skills. Providing targeted training and clear guidelines will empower volunteers to perform their roles effectively, ensuring that the institution's efforts are both inclusive and robust. The below points explore key insights relevant to mobilising and deploying volunteers to carry out the important task of scanning and assessing the virtual space.

— *Volunteers and Political Bias:*

It's crucial to harness the insight and engagement of volunteers who are *1) actively involved in political discourse* and who *2) follow key influencers*. Interestingly, it is precisely these politically engaged individuals—often viewed with suspicion due to potential biases—who can be the most discerning when it comes to evaluating the truthfulness of information[36]. Unlike swing voters, whose defining feature is their relative disengagement from politics, politically engaged individuals possess a deeper understanding of context, nuances, and the intricate details surrounding political narratives. This makes them far more adept at distinguishing between what is true and false. This insight challenges the common intuition that fact-checking should be done by apolitical or neutral individuals; instead, the goal should be to leverage politically engaged people in a way that ensures no single viewpoint dominates the process.

— *Quality Control and Diverse Perspectives:*

To ensure that contributions from these individuals are both accurate and balanced, we should implement techniques that have proven effective in crowdsourced environments. For example, requiring contributors to provide detailed explanations for their assessments not only helps filter out low-quality input but also encourages a culture of thoughtful, reasoned evaluation. Cross-validation, where multiple individuals independently assess the same content, ensures that the fact-checking process isn't hijacked by any one ideological group, maintaining a balanced and nuanced evaluation.

— *Innovative Incentive Structures:*

A key challenge nevertheless remains how to incentivise accuracy and discourage partisan flagging. Taking inspiration from prediction markets[37], a similar market-based approach offers a promising solution. By introducing a 'reputation currency' such as a simple rating score, volunteers could gain or lose score based on how often their assessments align with verified outcomes. This creates a self-regulating mechanism that encourages workers to make well-considered judgments, as they stand to gain or lose based on the quality of their assessments. Over time, this approach would cultivate a meritocratic environment where influence is earned through demonstrated accuracy.

— *Sensing a Network of Influence:*

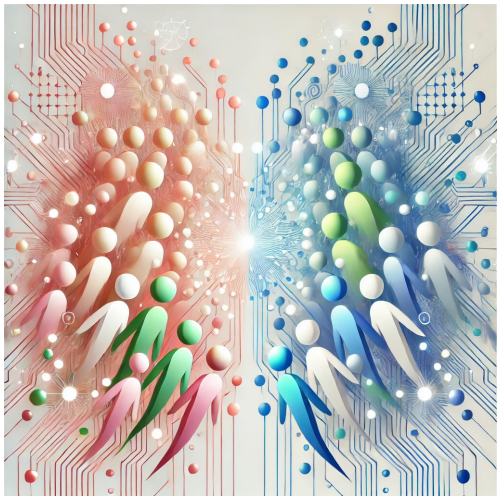
The real power of this system lies in building a diverse and balanced network of politically engaged individuals who follow key influencers across the ideological spectrum. This could be achieved by actively recruiting crowd workers who engage with different types of political content, ensuring representation from multiple viewpoints. For example, network analysis techniques could be used wherein social media help identify individuals who frequently interact with political influencers across various platforms, making them ideal candidates for fact-checking roles.

These key partnerships with platforms like X, Reddit, or Meta could help identify and recruit these individuals based on their engagement patterns, ensuring the pool of fact-checkers includes people who are not just politically engaged but also represent a wide array of perspectives - and that the network of influential voices is effectively monitored.

The most effective misinformation detection system doesn't shy away from politically engaged individuals; instead, it harnesses their deep understanding and commitment to uncover the truth. By carefully selecting, incentivising, and structuring a diverse network of political crowd workers, we can create a dynamic, self-regulating fact-checking ecosystem. This approach combines human insight, reputation-based incentives, and robust quality control mechanisms, offering a powerful solution to the challenges of misinformation in a politically complex landscape.

EXAMPLE: USHAHIDI'S VOLUNTEER MODEL

Ushahidi, a crowdsourcing platform developed to map reports of violence during the 2007 Kenyan elections, offers an effective volunteer recruitment model. By partnering with NGOs and community groups, Ushahidi recruits local volunteers who provide real-time, relevant data. This community-driven approach not only enhances data accuracy but also fosters a sense of ownership among volunteers, making the initiative more resilient and responsive. Electoral integrity institutions can adopt similar strategies, forming partnerships with local groups to build a robust, engaged volunteer network grounded in community involvement.



'Collective Power of Humans and Machines' (Generated by Dall-E)

HARNESSING THE COLLECTIVE POWER OF HUMANS AND MACHINES

As synthetic content increasingly dominates the digital landscape, relying solely on human volunteers to identify and manage misinformation is rapidly becoming unsustainable. Experts have predicted that soon the vast majority of online content could be synthetically generated[38], making it virtually impossible for human teams to keep pace without significant technological support. While human judgement remains crucial, its effectiveness is limited by the sheer volume and complexity of the content they must monitor.

In the past, human-driven efforts might have sufficed to counteract misinformation. However, with the exponential growth in synthetic media, the volume of content far outstrips human capacity. This surge necessitates the integration of cutting-edge technologies to automate the detection and analysis of potentially harmful synthetic content. These technologies do the heavy lifting, sifting through massive datasets to detect anomalies, behavioural patterns (of coordinated inauthentic behaviour, to use Meta's recent language) as well as suspicious content. This technological support allows human experts to focus their attention where it is most needed—on cases that require nuanced human judgment, ethical considerations, or complex decision-making.

Paradigm Shift: This is a Dissemination Problem

In the battle against synthetic media like deepfakes, considerable resources have been invested in developing detection tools. The academic community has produced a wealth of research on this topic (see systematic review[39]), exploring and refining detection methods. This has been mirrored by an intense focus within big tech and start-ups on finding the next breakthrough in identifying deepfakes.

However, the effectiveness of detection tools is increasingly questioned, especially as

synthetic content becomes more sophisticated. Text and audio, in particular, present significant challenges for detection methods, with current tools often falling short of accurately identifying AI-generated content in these formats. Video detection, while more advanced, is still not immune to these limitations, particularly as deepfakes become more realistic and harder to distinguish from genuine content. The adversarial nature of the GAN technology means that detection, whether performed by humans, AI, or a combination of both, is likely to remain a reactive rather than a proactive tool.

Academic literature suggests that some of the most promising detection efforts involve a combination of machine learning and human expertise, leveraging the strengths of both while mitigating their individual weaknesses[40]. However, given the challenges, especially with the anticipated rise of synthetic content, the limitations of detection become even more pronounced.

The expert consensus appears to be that without reliable digital provenance and watermarking techniques, consistently identifying synthetic content remains nearly impossible. Therefore, while detection remains a valuable tool, it is not sufficient to counter the threats posed by synthetic media.

Instead, we must turn our attention to tools, technologies, and methods that approach the mitigation of synthetic content harms as a dissemination problem rather than just a content problem. This paradigm shift is crucial as it unlocks more promising strategies for addressing the impact of synthetic media. By focusing on how harmful synthetic contents spread and the mechanisms through which they reach and influence large audiences, we can develop more effective mitigation strategies, avoiding the gridlock associated with monitoring and moderating vast volumes of content.

Tapping Into the Cutting-Edge

The Deepfake Detection System (DDS)[41] offers a cutting-edge approach that should be on the institution's radar as part of a broader strategy to combat synthetic content. This system stands out by combining decentralised deep-learning models with collective intelligence, all secured within a blockchain environment. This dual approach not only enhances detection accuracy but also ensures transparency and integrity throughout the process, addressing a key vulnerability in existing detection frameworks.

By harnessing blockchain technology, the DDS mitigates the risks associated with centralised control, making the detection process resistant to manipulation. The integration of user-driven voting mechanisms, weighted by reputation, alongside machine learning outputs, adds a layer of democratic validation that strengthens the reliability of detection results. As the architects of the electoral integrity institution consider their options, the DDS offers a compelling solution that goes beyond traditional detection. It aligns well with the paradigm shift toward focusing on the dissemination pathways of harmful content rather than just the content itself.

Early Warning Signals and Scanning Fringe Platforms

One of the most critical uses of AI in combating disinformation is its ability to detect early warning signals by continuously scanning fringe platforms like Telegram, 4chan, and Reddit. These less-regulated, smaller platforms often serve as breeding grounds for disinformation

campaigns and synthetic content. As[42] report in their recent white paper, AI tools can monitor these spaces in real-time, identifying emerging threats based on engagement patterns, recurrence, and other predictive analytics. By catching these signals early, institutions can take preemptive action, deploying countermeasures before harmful narratives have the chance to spread to mainstream platforms. This proactive approach is crucial in mitigating the impact of disinformation and maintaining the integrity of the information ecosystem.

Spotting Behavioural Patterns

In addition to early detection, Logically also show that AI technologies are adept at identifying behavioural patterns indicative of coordinated inauthentic behaviour (CIB). These patterns include synchronised activities such as coordinated posting, similar linguistic styles across accounts, and shared digital assets. By focusing on these behavioural cues, AI can uncover the hidden networks driving disinformation campaigns, allowing for timely intervention and disruption before these efforts reach a broader audience. This is a shift in thinking that the discipline cybersecurity experienced over a decade ago as it was adapting to botnet-related threats. It's important that misinformation follows suit.

Example of Bringing Humans and Machines Together: HAMLET

While there are numerous approaches to tackling misinformation, the HAMLET (Human and Machine in the Loop Evaluation and Training)[43] framework stands out as a strong example of what good looks like. This is not the only solution, but it effectively demonstrates how to blend AI with human expertise in the fight against disinformation, particularly in complex and dynamic environments.

— *Why is this important:* HAMLET exemplifies how human intelligence and machine learning can be combined to create a system that not only scales effectively but also maintains the necessary level of contextual understanding and ethical oversight. AI excels in processing large volumes of data, identifying patterns, and making preliminary classifications. However, the complexities and nuances of misinformation often require human input, especially in areas where context and ethical considerations are critical. HAMLET is designed to seamlessly integrate this human input, ensuring that the system remains both accurate and adaptive.

— *How it works:* The framework operates through a comprehensive workflow that includes mechanisms for collecting expert annotations, providing ongoing feedback, and monitoring the performance of AI models. This continuous interaction between AI and human experts ensures that the system evolves with the changing nature of online threats. HAMLET supports various data types, including text, speech, and multimedia, allowing it to tackle misinformation across multiple platforms and formats. By incorporating human oversight at key stages, the framework ensures that AI models are not only efficient but also fair, transparent, and aligned with ethical standards—critical for tasks like safeguarding electoral integrity.

— *Key strength:* One of HAMLET's strengths lies in its ability to detect coordinated inauthentic behaviour (CIB). Instead of focusing solely on individual pieces of content, HAMLET analyses broader behavioural patterns across networks,

such as synchronised posting or coordinated use of hashtags. This macro-level approach is essential in an environment where disinformation is becoming increasingly sophisticated and challenging to detect through traditional means.

— *Benefits and Challenges:* HAMLET's integration of AI and human expertise offers numerous advantages, such as scalability, adaptability, and the efficient handling of large data volumes. However, it also presents challenges, including ensuring the quality and consistency of human input, managing potential biases introduced by human contributors, and maintaining the transparency of AI models. To address these issues, HAMLET incorporates rigorous data quality management and frequent model updates, ensuring that the system remains effective and fair.

In the realm of electoral integrity, HAMLET's ability to merge large-scale data analysis with human oversight makes it particularly valuable. As the threats posed by synthetic content and disinformation continue to evolve, institutions need systems that can quickly adapt and integrate new insights. HAMLET's approach allows human experts to focus on the most complex and impactful cases while relying on AI to handle the more routine but large-scale tasks. This balance is crucial for maintaining the integrity of electoral processes in a digital world where both the tools and the tactics of misinformation are constantly advancing.

CASE STUDIES: PARTICIPATORY AND COLLABORATIVE SCANNING OF THE DIGITAL SPACE

COFACT - TAIWAN'S PARTICIPATORY FACT-CHECKING

Taiwan has been remarkably effective in countering misinformation, despite being one of the most geopolitically targeted countries in the world[44]. The success of Taiwan's approach can be partly attributed to Cofact, a collaborative fact-checking platform that operates across both social media and messaging apps like Line, which is widely used in Taiwan and functions similarly to WhatsApp.

Cofact leverages a crowdsourced model, where citizens are encouraged to report and verify information. When users encounter suspicious content, whether on social media platforms like Facebook or within private messaging apps, they can submit this content to Cofact for verification. The platform then cross-references the submitted content with its growing database of fact-checked information. If the content has already been verified, Cofact provides an immediate response to the user. If not, the content is escalated to professional fact-checkers and experts who evaluate its veracity.

Cofact's integration with Line is particularly significant, as it allows users to directly forward questionable content from private chats to the platform for verification. This functionality is critical given the private nature of messaging apps, where misinformation can spread unchecked and rapidly. By addressing misinformation on both public social media platforms and within the more closed environments of messaging apps, Cofact ensures a broader and more comprehensive approach to countering false information.

Cofact's model is built on a foundation of community involvement. It encourages ordinary citizens to participate in the verification process, fostering a sense of collective responsibility and increasing public trust in the fact-checking process. This participatory approach has been crucial in maintaining Taiwan's low levels of societal polarisation, even amidst relentless external misinformation campaigns.

LITHUANIA'S 'ELVES'

In Lithuania, persistent Russian disinformation campaigns gave rise to a unique civic movement known as the "Elves." This volunteer network emerged in response to the increasing influence of Russian troll farms and their attempts to spread false information and propaganda. Composed of ordinary citizens from various walks of life, the Elves aim to protect Lithuania's information ecosystem by identifying disinformation, flagging fake accounts, and alerting the public to the dangers of foreign influence online. The group, which counts as many as 22,000 members, operates primarily by scanning social media platforms for misleading content and working to expose fake news before it gains traction.

The Elves' impact is significant, and their efforts are widely acknowledged, even by official

channels. Lithuania's Ministry of Defence has described their actions as "very helpful in exposing damaging propaganda."[45] The Elves work to ensure that harmful content linked to foreign state actors, particularly from Russia, is swiftly reported and often removed by social media platforms. This grassroots effort complements the official governmental strategies for countering disinformation, providing an additional layer of resilience against external threats.

Operating in their spare time, the Elves serve as a decentralised and agile force, highly effective in adapting to the fast-evolving landscape of online disinformation. Their collective vigilance plays a key role in preserving the integrity of public discourse in Lithuania, acting as a civilian counterforce to Russia's ongoing efforts to undermine democratic stability in the region.

TSEK.PH AND CEKFAKTA - COLLABORATIVE FACT-CHECKING IN SOUTHEAST ASIA

Tsek.ph in the Philippines and CekFakta in Indonesia represent collaborative fact-checking models that have proven effective, particularly in the context of elections. Both platforms bring together a consortium of media organisations, civil society groups, and academic institutions to coordinate their fact-checking efforts. This collaboration is designed to scan the digital space more effectively, streamline the verification process, reduce duplication of efforts, and ensure a more efficient response to the spread of misinformation.

In both the Philippines and Indonesia, elections have been marred by the proliferation of misinformation, often designed to influence voter behavior or discredit political candidates. Tsek.ph and CekFakta address this challenge by pooling the resources and expertise of various stakeholders. Each participating organisation contributes to a shared database of fact-checked content, which is accessible to all members of the consortium. This shared approach not only conserves resources but also ensures that false information is quickly identified and debunked across multiple platforms.

A key strength of Tsek.ph and CekFakta is their ability to prevent the redundancy that often plagues independent fact-checking initiatives. By collaborating, these organisations avoid the inefficiencies of verifying the same content multiple times. This is particularly important in the fast-paced information environment surrounding elections, where the timely correction of misinformation is critical.

Moreover, the collaboration between media, civil society, and academia in these initiatives fosters a more comprehensive and multifaceted approach to fact-checking. It combines the immediacy and reach of media organisations with the analytical depth of academic research and the grassroots connections of civil society. This synergy enhances the credibility of the fact-checking process and strengthens public trust in the information being disseminated

Lessons For an Electoral Integrity Institution

- **Cross-Platform Engagement:** As seen in Cofact's integration with social media and messaging apps, addressing misinformation requires a multi-platform strategy. Electoral integrity institutions should consider similar approaches, ensuring that fact-checking capabilities are accessible on both public platforms and private communication channels.

- **Collaborative Models:** The success of Tsek.ph and CekFakta underscores the importance of collaboration among different stakeholders. Institutions should foster partnerships between media, civil society, and academia to create a unified front against misinformation, reducing redundancy and enhancing the overall impact of fact-checking efforts.

- **Community Involvement:** Engaging the public in the fact-checking process, as demonstrated by Cofact, can build trust and enhance the effectiveness of misinformation countermeasures. Electoral integrity institutions should explore ways to involve citizens in reporting and verifying information, thereby creating a more resilient information environment.

ASSESS EFFECTIVELY AND IMPARTIALLY

In an era where synthetic misinformation poses significant risks to public discourse and electoral integrity, institutions must develop sophisticated strategies for triaging content identified through digital scans. Triaging involves categorising content based on its potential threat and determining the appropriate level of intervention. This chapter outlines a structured approach to triaging that combines automated tools with human oversight, emphasising the need for clear guidelines, timely decision-making, and transparency.

In this section, we concentrate on two key aspects of assessment:

1. Triaging *Effectively*
2. Triaging *Ethically and Impartially*

TRIAGING EFFECTIVELY

Establishing Clear Guidelines and Criteria:

A critical component of effective triaging is the establishment of clear, actionable guidelines that dictate how different types of content should be handled. These guidelines should be informed by legal frameworks, ethical considerations, and the specific mission of the institution.

These guidelines must be comprehensive, providing clear criteria for when content should be frozen, labeled, or left alone. For instance, content that poses an imminent threat to the

electoral integrity such as fraudulent information about the polling location may require immediate action, whereas misleading but non-threatening information about one of the candidates might only need a label or added context. By clearly defining these criteria, institutions can ensure consistency and fairness in their responses, minimising the risk of overreach or neglect.

To effectively triage content in the digital information environment, institutions must develop and apply clear criteria for categorising content based on its potential risk and the necessary response. This involves assessing the harm potential, reach and influence, and the credibility of the content's source.

- *Harm Potential:* The first consideration is the potential harm posed by the content. This includes evaluating whether the content is likely to incite violence, spread dangerous misinformation, or undermine public trust. For example, content that promotes electoral fraud or violence would necessitate immediate intervention, while a misleading claim about a candidate might only require labeling rather than removal.

- *Falseness:* This involves determining whether the content is outright fabricated, manipulated, selectively edited, partially true but misleading, or simply out of context. An outright fabrication due to its potential to entirely distort reality may require a different response as compared to an out-of-context information. The institution should consider the potential impact and the intent behind such falsifications, differentiating between satire, unintentional error, and deliberate deceit.

- *Reach and Influence:* The reach of the content is another critical factor. Content that is going viral or being shared by influential figures or platforms may have a greater impact on public perception and, therefore, should be prioritised for review. High-reach content, even if only moderately harmful, can have significant consequences due to its wide dissemination.

- *Virality Potential:* Evaluating virality potential means identifying elements that could fuel rapid dissemination, such as emotional triggers, polarising narratives, or alignment with pre-existing biases in society. Highly sensational or politicised content, particularly on controversial topics, often spreads exponentially across platforms. By using predictive algorithms, sentiment analysis, and human judgement, the institution can pre-emptively identify content with a high likelihood of virality, prioritising swift action on pieces likely to reach a critical mass quickly, thereby minimising potential harm before it reaches its peak influence

- *Source Credibility:* Finally, the credibility and intent of the content source should be analysed. Content from known disinformation actors or foreign malicious sources might require stronger actions, such as suppression or removal, compared to content from local, credible but misguided sources, which might only need correction or additional context.

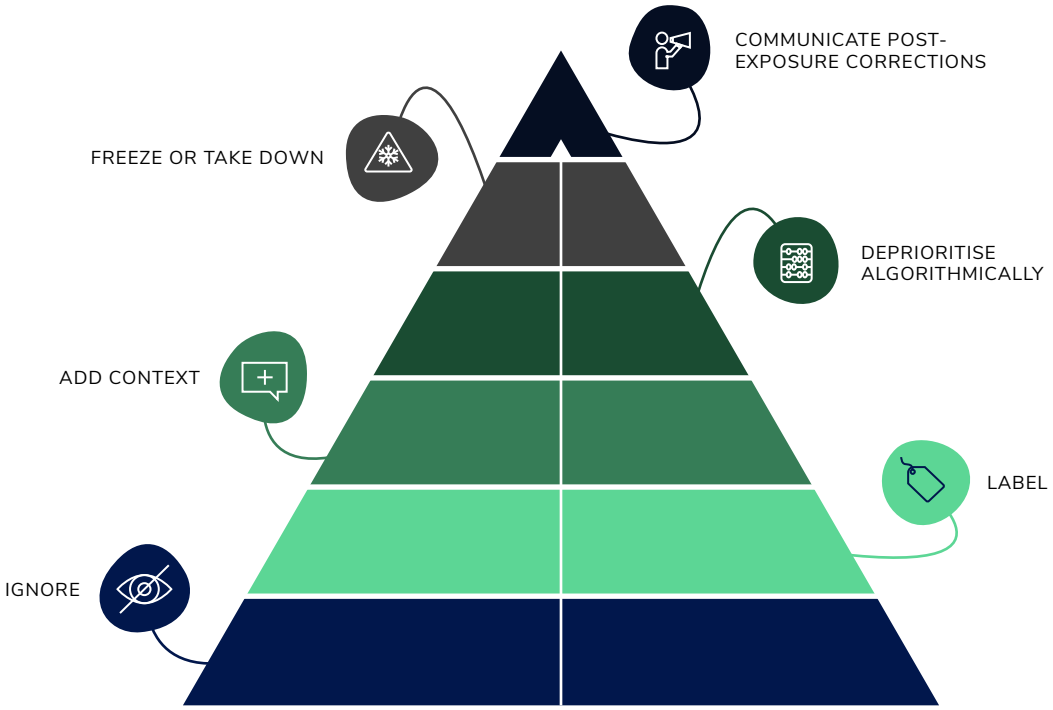
These guidelines should be rigorously tested through "red-teaming" exercises, where hypothetical scenarios are used to stress-test the decision-making process and ensure that

responses are both effective and legally sound.

Developing a Tiered Response System

A tiered response system is essential for effectively managing digital content based on its risk and potential impact. This system categorises content into different levels of severity, allowing for responses that avoid overreach while addressing the most serious threats effectively.

Figure 2. Tiered Response System: scaling intervention based on harm and risk



1. **Ignore or Monitor:** Some content, while false or misleading, may have limited reach or negligible impact. In such cases, monitoring rather than intervening can be more effective. Over-policing can lead to accusations of censorship or trigger the Streisand effect, where attempts to suppress information inadvertently amplify its spread. By monitoring instead of acting, institutions can keep a watchful eye on potential escalation without unnecessary intervention.

2. **Label:** For misleading content that does not pose an immediate threat but could still misinform, applying warning labels is an effective strategy. This approach educates the public without infringing on free speech. Platforms like Twitter and Facebook have used this strategy by flagging content as disputed or linking it to credible sources. Additionally, there is an increasing need to label content as AI-generated when applicable. This transparency helps users assess the credibility of the information and understand the origin of the content.

3. **Add Context:** Alongside labeling, it is also important to provide additional context to help users form a balanced view. For instance, services like Ground News offer a feature that shows how different news outlets cover the same story, helping to highlight biases and provide a more comprehensive perspective. Adding context ensures that users are not just informed that content is disputed but are also given the tools to understand the broader picture.

4. **Deprioritise or Suppress:** Content that is considered harmful but not illegal can in certain cases be de-amplified by algorithms to reduce its visibility. This tactic allows the content to remain accessible without giving it undue prominence, thus mitigating its potential harm. This approach balances the need to address harmful content with the principle of free expression. Content that is harmful yet protected under free speech laws may be deprioritised algorithmically, limiting its reach while avoiding outright censorship.

5. **Freeze or Take Down:** In cases where content is deemed highly dangerous, such as incitements to violence or misinformation with serious public health implications, immediate freezing or removal may be necessary. These actions should be taken in accordance with legal frameworks and platform policies to avoid overreach and ensure that responses are appropriate and justified. Additionally, when content is taken down, it may be necessary to reach out to individuals who viewed or engaged with the content to inform them that it was false. Although studies show that correcting misinformation does not always change opinions, it remains a critical step in combating the spread of false information and maintaining public trust.

6. **Communicate Post-Exposure Corrections:** If individuals encounter false content, it's crucial they are retrospectively informed that it was false. Research supports that while notifying individuals may not always correct false beliefs, it still plays a role in combating misinformation. For instance, a meta analysis found that simple corrections can reduce the perceived credibility of false information, but these corrections are less effective for emotionally charged misinformation that aligns with people's beliefs[46]. Therefore, while not foolproof, such notifications remain essential.

Research indicates that correcting misinformation after exposure is not always highly effective, as belief perseverance bias (BPB) can cause individuals to maintain their initial false beliefs despite the correction. Techniques like "counter-speech" (presenting arguments against the misinformation) and "awareness training" (raising awareness of BPB) have proven more effective than simply providing factual corrections[47]. Incorporating these methods into the communications would thus be important to increase the effectiveness.

Leveraging Automation and Human Expertise

Effective triaging begins with the integration of advanced automated tools alongside human expertise. Automation is crucial for managing the scale of digital content, particularly in the initial stages of assessment. Tools that transcribe video content and analyse it through fact-checking algorithms can categorise content by potential risk levels, allowing human analysts to focus on the most pressing cases. As organisations on the cutting-edge of the misinformation sector[48] continue to develop these tools, they are becoming increasingly adept at pre-triaging content, identifying potential risks with greater accuracy and efficiency. While machines are invaluable for filtering vast amounts of data, the final decision-making process must incorporate human judgment to account for the nuanced understanding that AI might lack.

Implementing Rapid Response Teams

For content that poses immediate risks, such as potential incitements to violence or critical misinformation during election periods, rapid response teams are essential. These teams, composed of experts in law, communications, technology, and public policy, must be prepared to act swiftly, using the pre-established guidelines and decision trees to determine the best course of action.

The effectiveness of these teams hinges on their ability to operate with both autonomy and accountability, ensuring that actions taken are in line with the institution's overarching mission and legal frameworks.

Bringing the Crowds Back in

In conjunction with automated tools and expert analysis, the institutions can leverage the wisdom of crowds at various stages of the assessment process. This may involve gathering diverse perspectives from a broad audience to assess the content in question in scenarios when the expert teams fail to make progress. However, the need for prompt action must be carefully balanced against the time required to collect and analyse these opinions. In time-sensitive situations, waiting for extensive crowd feedback may not be feasible, and institutions must have mechanisms in place to make timely decisions.

To effectively harness crowd wisdom, institutions should start with a small, diverse group for initial triangulation. If the content remains unresolved, the inquiry can be broadened to include more voices.

How Assessment Might Work in Practice

The triaging system would integrate automated tools, volunteers, and rapid response teams:

1. *Initial Automated Screening:* Automated systems, as described earlier, act as the first line of defence, scanning content in real-time. Depending on the automated assessment's verdict, content deemed lower-risk is sent to crowd workers for further review, while high-risk content is escalated directly to the rapid response teams.

2. *Crowd Evaluation:* Crowd workers act as a corrective and calibrating mechanism for the automated tools. They provide context, assess the nuances missed by algorithms, and ensure accuracy. Certain content is double-checked by both the automated systems and human evaluators, reinforcing precision. Crowd workers handle lower-priority tasks like labelling or adding context, ensuring that rapid response teams are only tasked with high-stakes issues.

3. *Rapid Response Team Decision-Making:* High-risk content is handled by expert rapid response teams, who review the data and insights provided by automated tools and crowd workers. These teams make recommendations to the authority empowered to act—deciding whether content should be frozen or taken down, but not dealing with lower-priority tasks like labelling content as AI-generated.

4. *Communication and Transparency:* Once decisions are made, transparent communication is essential. Individuals who interacted with or viewed the false content would be retrospectively informed about its inaccuracy (detailed below), maintaining trust and accountability in the institution's process.

TRIAGING ETHICALLY AND IMPARTIALLY

Impartiality in triaging is crucial for the institution because it ensures that actions taken against content are based solely on objective criteria rather than subjective biases, power imbalances, or external pressures. This impartial approach not only upholds the ethos and integrity of the institution - it also fosters public trust, which is essential for maintaining the legitimacy of its actions. When triaging is done impartially, it prevents any perception or reality of censorship being driven by political, ideological, or commercial interests, thereby protecting the democratic process and the principle of free expression. By committing to impartiality, the institution can effectively navigate the complexities of content moderation, ensuring that interventions are justified, proportionate, and consistent with legal and ethical standards.

Incorporating Legal Frameworks and Ethical Considerations

All triaging actions must be grounded in existing legal frameworks and ethical considerations. This involves ensuring that any intervention aligns with national laws, human rights standards, and platform policies. Institutions must balance the need for effective intervention with the protection of free speech and privacy rights, ensuring that actions are proportionate and transparent.

Clear guidelines should be established for when it is acceptable to freeze or remove content. This means that the institution needs to ask its stakeholders serious searching questions around what is and isn't ethical. An illustrative example may occur during critical periods when media is restricted from reporting on elections - commonly referred to as the "media blackout" or "election silence". This period typically occurs right before or during an election to prevent the influence of last-minute campaign coverage on voters. Given that no such rule applies in the world of social media, would it be ethically to turn the dial up (and possibly run the risk of overshooting censorship) on censoring content that's identified as harmful, in the interest of protecting the integrity of elections?

Transparency and Public Communication

Transparency is a cornerstone of public trust and institutional credibility. Institutions must be proactive in communicating their triaging processes and decisions to the public. This includes explaining the criteria used for categorising content, the steps taken in response, and the reasons behind those decisions. Going the extra mile in transparency not only mitigates potential backlash but also reinforces the institution's commitment to fairness and accountability.

Ensuring the power to act is granted within a legally accountable and transparent framework is essential. Any institution charged with combating electoral disinformation must operate with clear legal oversight to prevent overreach and protect civil liberties. Transparency in its operations can enhance public trust and keep the institution accountable to democratic principles. Publishing transparency reports, documenting decisions, and openly sharing criteria for intervention are critical measures that not only ensure accountability but also strengthen the institution's credibility.

Inspiration can be drawn from radical transparency models, such as those employed by g0v in Taiwan, serve as instructive examples. As an intentional design feature, this civic tech movement keeps all meeting minutes publicly available, making every decision open to scrutiny. This type of openness ensures that the institution is continuously held accountable to both its stakeholders and the public. Such transparency also serves to depoliticise sensitive decisions around content moderation by providing a factual basis for public debate.

CASE STUDY: A CAUTIONARY TALE (GDI)

The Global Disinformation Index (GDI), an organisation dedicated to assessing and ranking media outlets based on their risk of spreading disinformation, recently found itself embroiled in controversy. The GDI faced widespread criticism for allegedly allowing ideological biases to influence its ratings, rather than relying solely on objective assessments. This scrutiny extended beyond concerns about bias, focusing on GDI's failure to maintain transparency—one of its core values. Critics argued that the organisation's decision-making processes were very opaque, with little to no explanation provided when transparency was requested. This lack of openness not only eroded trust in the GDI's rankings but also raised questions about its operational integrity.

The situation escalated to the point where the then British Foreign Secretary David Cameron publicly announced that the UK Government would no longer fund the Global Disinformation Index. This decision underscored the critical importance of impartiality and transparency, particularly for organisations that hold significant sway over public perception and policymaking. The GDI's case serves as a stark reminder that, in the battle against disinformation, the tools and institutions we rely on must themselves be above reproach, free from ideological influence, and committed to the highest standards of transparency.

Embedding Impartiality – Lessons from ICC and BBC

To design an electoral integrity institution with a foundation of impartiality, we can draw on the principles and mechanisms embedded in well-established global institutions known for their objectivity and credibility. Two such examples are the International Criminal Court (ICC) and the British Broadcasting Corporation (BBC), each of which incorporates distinct design features that ensure impartiality.

The **International Criminal Court (ICC)**, based in The Hague, provides a robust example of how impartiality can be embedded into the fabric of an institution. The ICC is governed by the Rome Statute, which was negotiated by a broad coalition of countries, ensuring that no single nation or group of nations could dominate its agenda. The judges and prosecutors of the ICC are elected by the Assembly of States Parties, which consists of representatives from all the member countries, providing a diverse and balanced oversight. Moreover, the election process is designed to prevent any concentration of power, with terms of office limited and geographical and gender diversity mandated. This inclusive and democratic approach helps to safeguard the court's impartiality, ensuring that it operates above national politics and is viewed as a neutral arbiter of international law.

The **British Broadcasting Corporation (BBC)** offers another excellent model of institutional impartiality. As the UK's national broadcaster, the BBC has a mandate to provide unbiased and accurate information to the public. This mandate is safeguarded by its Royal Charter, which enshrines the BBC's independence from government and commercial pressures. The BBC is overseen by the BBC Board, which is appointed through a process designed to ensure a balance of perspectives and is subject to public scrutiny. Editorial decisions within the BBC are guided by strict internal guidelines that emphasise impartiality, accuracy, and fairness. The organisation's commitment to transparency is further demonstrated by its regular publication of editorial standards and its willingness to undergo external audits and reviews. These mechanisms collectively help the BBC maintain its reputation as a trusted and impartial news source, free from undue influence.

CASE STUDY: BALANCING SECURITY AND FREEDOM (PDA)

The Swedish Psychological Defence Agency (PDA) represents a sophisticated model of a modern democracy's defence against foreign disinformation, operating with a profound commitment to safeguarding civil liberties. Officially re-established in January 2022, under the Ministry of Defence, the PDA is the culmination of Sweden's 70-year history of psychological

defence, an institution that originated during the Cold War. Despite its roots in a bygone era, the agency has evolved to address contemporary challenges, particularly the threat posed by foreign malign information influences that target Sweden or its national interests.

One of the PDA's most striking features is its clear and deliberate separation between countering external threats and safeguarding the rights of its citizens. The agency's mandate is confined to dealing with foreign disinformation that is antagonistic, deceptive, and intended to harm Sweden. This focused approach is crucial in maintaining the agency's impartiality and protecting it from potential misuse as a domestic censorship tool. Swedish citizens, even if they propagate false information, are not treated as threats but rather as 'vulnerabilities' who might need support to build resilience against foreign disinformation. This crucial design choice ensures that the PDA operates within a framework that respects and upholds the democratic principles of free speech and civil rights.

In its operations, the PDA employs a carefully calibrated approach to countering disinformation. Rather than immediately resorting to censorship or overt confrontation, the agency first seeks to strengthen public awareness and understanding, equipping citizens with the knowledge to critically assess information themselves. This approach underscores the agency's commitment to not overstepping its mandate or infringing on freedoms guaranteed to the public. When necessary, the PDA escalates its response in measured steps, from contextualising and labelling questionable content to directly addressing and exposing the foreign actors responsible for the disinformation. This tiered strategy ensures that any intervention is proportionate, targeted, and transparent, thus maintaining public trust and upholding Sweden's democratic values.

The Swedish PDA serves as an exemplary case of how a state can protect itself from external disinformation while ensuring that its operations do not infringe on the freedoms and rights of its citizens. The agency's structure and methods provide valuable lessons in balancing national security with the preservation of democratic principles, making it a model for other nations facing similar threats to get inspired by.

ACT WITH POWER AND ACCOUNTABILITY

In this section, we:

- Discuss why it is critical to keep capability and power separate, but closely linked.
- Discuss what can be done about the challenges related to politically appointed roles.
- Present an example governance model — one that balances power, speed, and accountability.

THE GAP BETWEEN POWER AND CAPABILITY

When establishing an electoral integrity institution, the power to act is paramount. It is not enough to merely have the capability to monitor disinformation, identify foreign interference, or spot synthetic content. Without the authority to execute corrective actions, such as freezing disinformation, taking it down, or disseminating counter-information, these capabilities amount to little more than academic exercises. The institution must have both the mandate and legal backing to take swift, decisive action. Without this power, the institution risks being reactive and unable to mitigate the rapid harm that modern disinformation campaigns can cause, especially during sensitive electoral periods.

The case of Keir Starmer's deepfake video highlights this critical point. The video, which falsely depicted the British politician making damaging statements, was identified as disinformation by experts and even rival politicians[49]. Yet, despite this being publicly acknowledged, X (formerly Twitter) refused to take it down[50]. This situation underlines the critical need for executional power—the ability not just to identify, but also to act on harmful content before it spreads unchecked. If platforms or political actors can refuse or delay removal of disinformation, it renders an integrity institution's capability to detect these threats effectively futile without enforcement powers (see *Appendix C* for a polemic on the ethics of pre-monitoring content during electoral period)

For such an institution to wield power effectively, buy-in from key stakeholders such as governments, social media platforms, and the general public is critical. Social media platforms play a significant role, as they control much of the digital space where disinformation thrives. Political will may be essential to bind these platforms to cooperate. Additionally, public support for the institution's efforts, built on transparency and trust, can apply pressure on platforms to act in alignment with the institution's recommendations. A well-informed public that understands the harm of disinformation is more likely to support and trust the institution's actions, creating a reinforcing loop of legitimacy.

Lessons from VIGINUM: Separating Power from Capability

A lesson can be drawn from France's VIGINUM, an example of an existing institution we give above. Attached to the General Secretariat of Defense and National Security, the agency identifies large-scale disinformation campaigns, particularly those involving foreign actors, but the responsibility for correcting or removing disinformation lies with democratic institutions. As VIGINUM's director Marc-Antoine Brillant emphasised, it is up to politicians, the media, and eventually the courts of justice to correct "untruths"[51]. For the sake of democracy, he said, this function of "correction" should be carried out by democratic institutions, not by a government agency. This separation is essential in ensuring that government agencies do not control public discourse and that any action taken against disinformation has democratic legitimacy. VIGINUM's role as a monitoring entity underscores the importance of detecting threats without overreaching into enforcement—a principle that should guide the design of any electoral integrity body.

NAVIGATING POLITICAL APPOINTMENTS TO ENSURE ACCOUNTABILITY

In many governance contexts, political appointments are not only common but may also be unavoidable. They can ensure institutions remain connected to broader democratic structures

and reflect diverse political perspectives. However, it is essential to design mechanisms that balance the benefits of political representation with safeguards against undue influence or perceptions of bias.

One effective approach is to implement **cross-partisan confirmation processes**, where appointees are vetted and approved by committees representing a range of political parties. For example, in the United States, Senate confirmation hearings for key roles provide transparency and enable scrutiny across political lines. Adopting a similar mechanism could help ensure that appointments to the institution are broadly acceptable and less likely to be perceived as partisan.

Another strategy is to establish **staggered terms for leadership positions**, similar to the model used by central banks like the Federal Reserve. With terms overlapping across election cycles, this structure prevents any single government from fully shaping the institution's leadership, promoting continuity and reducing the risk of partisanship dominating decision-making.

Lastly, **transparency in the appointment process** is key. Publishing detailed criteria for roles, documenting the selection process, and openly communicating the rationale behind appointments can help mitigate concerns about bias. Regular public reporting on the institution's governance practices further reinforces accountability.

Figure 3. Example Governance Model: Ensuring executive integrity through oversight and accountability

EXAMPLE GOVERNANCE MODEL: BALANCING POWER, SPEED, AND ACCOUNTABILITY

The effectiveness of this institution hinges not only on the capability of detecting and triaging synthetic content but - perhaps even more importantly - on possessing the decisive power to act against disinformation. Power must extend beyond individual seats or positions; it should be viewed as an interconnected network, enabling the entire system to respond swiftly rather than creating bottlenecks. When power is embedded across the system, it fosters resilience and ensures that interventions against disinformation are both swift and democratically accountable. The following features explore how power can be structured to balance pace with accountability:

— **Independent Electoral Commission as a Core Feature:**

An independent electoral commission, like those in the UK or India, serves as a natural 'home' for the power to act against disinformation, providing a centralised yet impartial authority. These commissions are well-positioned to wield this power effectively, as they combine the necessary legal mandate, expertise, and insulation from political pressures, making them the most suitable entities to take decisive, unbiased action when needed. Empowering this commission with clear, legally mandated authority would allow it to act decisively. Regular oversight from parliamentary committees, judicial review, and public transparency reports ensure that its actions remain accountable. This feature ensures that an independent body can act rapidly, especially during election periods, without being swayed by changing political winds.

— **Judicial Oversight and Emergency Powers:**

Empowering the judicial system with rapid oversight may be necessary in some scenarios to ensure that actions taken are both timely and legally legitimate. Specialised courts could handle cases swiftly. A recent example was seen in the UK's response to the Southport riots in 2024, where within a matter of days, hundreds of people appeared in courts on charges related to the disorder, and many were sentenced promptly[52]. This demonstrates how a judicial system can efficiently manage high-volume, urgent cases, providing a strong precedent for rapid interventions in disinformation threats during critical periods.

The judicial system should focus on high-impact cases such as attempts to incite violence, large-scale electoral interference, or coordinated disinformation campaigns that threaten public safety or democratic processes. These cases may require heightened legal scrutiny to ensure swift, authoritative decisions. Conversely, the system should avoid being bogged down by minor infractions, everyday falsehoods, or content that can be effectively handled by administrative bodies or independent commissions. The goal is to prioritise cases where the legal system's authority is crucial to maintaining order and trust.

— **Multistakeholder Board with Voting Mechanisms:**

A multistakeholder board can provide a democratic layer of oversight, incorporating voices from government, civil society, media, academia, and technology platforms. To avoid delays often associated with consensus-seeking, the board could employ cutting-edge voting mechanisms, such as "quadratic

voting," where stakeholders allocate votes based on the intensity of their preferences[53]. This technique allows for nuanced, balanced decisions, ensuring that all voices are considered without stalling the process. For particularly contentious issues where consensus isn't emerging, rapid deliberation could be introduced, similar to Taiwan's g0v movement, but only as a last resort to avoid unnecessary delays. Independent audits and transparent reporting would further enhance the legitimacy and accountability of this board.

A Hybrid Model in Practice

The most effective governance structure would most likely blend some or all of these features into a sophisticated hybrid model, combining the strengths of an empowered commission, judicial oversight, and multistakeholder involvement. Here's how this could work in practice:

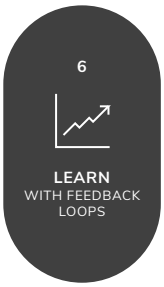
— *Primary Decision-Making:* Upon receiving the recommendations from the electoral integrity institution this paper fleshes out (i.e., the capability wing), the independent electoral commission would be granted the power to take immediate action on a majority of disinformation threats, especially during critical periods such as the lead-up to elections. This authority would enable rapid responses, such as issuing corrections or freezing content, without the delays typically associated with larger bureaucratic processes.

— *Immediate Review Mechanism:* Actions that are appealed or challenged would undergo judicial scrutiny. A specialised electoral court would handle appeals swiftly (e.g., within 24 to 48 hours), ensuring that interventions are legally sound and proportionate. This process creates an effective safeguard against potential misuse of power while allowing the commission to act decisively, maintaining a balance between swift action and legal accountability.

— *Multistakeholder Validation:* Alongside judicial oversight, a multistakeholder board would serve as an additional layer of scrutiny, providing feedback on decisions taken by the commission. This board would leverage advanced voting techniques to quickly provide input or recommendations, particularly on high-profile cases where public trust is at stake. For instance, if the commission were to delay or freeze highly politicised, inflammatory content, the board's rapid voting mechanism would validate or challenge this action, ensuring diverse viewpoints are considered.

— *Emergency Powers and Sunset Clauses:* The commission could be granted temporary emergency powers during critical electoral periods, such as the last month before an election. However, these powers would be subject to sunset clauses, meaning they expire after a set period unless renewed by a judicial or parliamentary body, ensuring that extraordinary measures remain temporary and proportionate.

— *Transparency and Public Accountability:* The entire process would be transparent, with all actions, voting records, and decisions published in real time or with minimal delay, ensuring public awareness and maintaining democratic legitimacy. Regular audits and reports would be mandated to assess the system's effectiveness and ensure all parties are held accountable.



LEARN VIA FEEDBACK LOOPS

To counter disinformation effectively, an electoral integrity institution must function as a sophisticated, learning organisation that evolves continuously in response to new challenges. Disinformation adapts, shifts, and manifests in increasingly complex ways, making it crucial for such an institution to remain dynamic, incorporating adaptive learning, cross-border collaboration, and innovative problem-solving to stay ahead. Below are some key principles to consider in order to design effective feedback loops.

Continuous Real-Time Learning and the Outside-In Approach

The electoral integrity institution should leverage advanced real-time data analytics, machine learning, and predictive modelling. However, it's not enough to rely solely on internal data. The institution must adopt an outside-in approach[54], integrating external perspectives and expertise. This involves engaging with external stakeholders—tech companies, civil society groups, academic institutions, and the public—through structured feedback mechanisms, public consultations, and data-sharing partnerships. By incorporating these external inputs, the institution stays attuned to emerging threats and can rapidly adapt its strategies. For example, tech companies can provide real-time insights on disinformation trends and manipulation tactics, keeping the institution's strategies timely and effective.

For instance, engaging directly with tech companies enables access to datasets on disinformation patterns, while partnerships with academic institutions offer cutting-edge research findings. By tapping into these external resources, the institution can adapt its strategies and stay ahead of evolving threats. This outside-in method ensures that the institution's strategies are informed, agile, and responsive to the broader disinformation landscape.

In practice, the institution should establish regular, structured dialogue and collaboration with its partners. Each electoral cycle should end with comprehensive debriefs and analysis, where the institution reviews successes and failures and integrates lessons learned into future strategies. Cross-sectoral learning isn't a static process but an ongoing engagement that allows the institution to stay relevant and agile.

The Threats Will Evolve: Designing for Adaptiveness

'Adaptiveness' (Generated by Dall-E)

Effective institutions must embrace continuous experimentation and flexibility, ensuring they can adapt to changing realities. Malicious actors are constantly refining their tactics, making it essential for the institution to remain adaptable and avoid overconcentrating on any single subset of threats. A recent example highlights the shift from automated bots to more sophisticated hybrid methods involving real individuals. The Tenet Media case[55], highlighted by the U.S. Department of Justice, alleges that Russian operatives covertly paid U.S. influencers to unknowingly spread disinformation framed as legitimate content. By bypassing traditional bot-driven strategies, this method exploited the credibility of real individuals to deliver messages with higher authenticity and reach.

Such tactics blur the lines between legitimate speech and disinformation, making detection and mitigation increasingly complex. To counter these hybrid approaches, the institution must maintain the flexibility to address a broad range of potential harms. This requires adopting adaptable threat detection methods, rather than focusing narrowly on a predetermined set of challenges. The ability to pivot and respond to evolving threat vectors is a critical defining feature of this institution, ensuring it can keep pace with the ever-creative tactics of malicious actors.

The Financial Stability Board (FSB) is a good example, as it deliberately avoided over-specifying its remit, allowing it to evolve and respond to emerging financial risks. In contrast, the EU's AI Act has been criticised for being overly prescriptive too early, making it difficult to adjust to the fast-evolving AI landscape[56]. Some have argued that the Act's detailed and prescriptive approach makes it challenging to assess systemic risks, as it doesn't account for interactions between AI models and other platforms, apps, or plugins[57]. By setting rigid guidelines too soon, it risks becoming outdated before implementation.

A more adaptive approach is exemplified by California's SB-1047 AI Bill, which purposefully left space for growth. It avoids over-specifying the rules for AI technologies and includes flexible provisions to keep up with rapid advancements. In other words, it is an example of concentrating on accountability, rather than clarity. The bill focuses on establishing guidelines and requires technology developers to integrate safeguards as they develop and deploy powerful, frontier models, setting a foundational framework while leaving room to adapt based on future developments and risks identified through real-world use. This flexibility allows the regulatory framework to evolve in line with technological shifts, ensuring it remains effective over time.

Global Real-Time Knowledge Sharing and Learning Networks

Disinformation is a global issue, and no institution can tackle it in isolation. Electoral integrity institutions should participate in a global learning network where knowledge is exchanged in real time. This network would allow institutions from different countries to share insights, methodologies, and threat intelligence instantly. These models have been emerging. One example of such a network is the Global Health Security Agenda (GHSA), which connects over 70 countries, international organisations, and private sector companies to rapidly share public health information, allowing for a coordinated global response to infectious disease outbreaks. We could draw inspiration from such a network to accelerate knowledge sharing, allowing electoral integrity institutions to respond faster and more effectively to emerging disinformation threats.

Peer review mechanisms, akin to scientific research, could be embedded in this network, too, allowing institutions to assess and improve each other's strategies, ensuring that the best practices are applied globally. By subjecting their methodologies, tactics, and outcomes to scrutiny from global counterparts, electoral integrity institutions can ensure that they are held to the highest standards. This benchmarking process not only fosters a culture of transparency and accountability but also ensures that the institution's strategies remain cutting-edge and grounded in the collective wisdom of international experts.

Dedicated Foresight Unit - Anticipation and Stress Testing

To effectively counter evolving disinformation threats, the institution must integrate a diverse range of strategic foresight methods into its framework. Techniques such as horizon scanning, open-source intelligence (OSINT), and red and blue teaming exercises provide a robust toolkit for identifying vulnerabilities, anticipating future challenges, and stress-testing defences.

'Foresight' (Generated by Dall-E)

Horizon scanning systematically tracks emerging trends, weak signals, and disruptive developments across technological, political, and societal domains. For example, it enables the institution to identify advances in generative AI or shifts in disinformation tactics before they escalate into significant threats. Complementing this, OSINT leverages publicly available data—such as social media activity, news reports, and online forums—to monitor disinformation patterns in real time. OSINT enables the institution to uncover early indicators of coordinated campaigns and provides actionable intelligence for rapid response.

Within this foresight ecosystem, red and blue teaming exercises remain indispensable. Red teams simulate adversarial tactics, probing for weaknesses in detection tools, protocols, and broader processes, while blue teams defend against these scenarios, testing and refining the institution's responses under

realistic conditions. These exercises foster a proactive stance, enabling the institution to develop defences that anticipate the likely moves of malicious actors.

To maximise the effectiveness of these methods, the institution must involve cross-disciplinary expertise spanning AI ethics, behavioural science, geopolitics, and cybersecurity. Collaborating with external stakeholders, such as technology platforms and civil society organisations, enhances the credibility and utility of insights. Furthermore, foresight outputs—from horizon scanning trends to OSINT intelligence and red teaming vulnerabilities—must feed directly into strategic oversight. This integration ensures that insights inform the development of detection technologies, prioritisation of mitigation strategies, and coordination with international partners.

A **dedicated Foresight Unit is essential** to institutionalise these activities. Reporting directly to the strategic oversight function, this unit would oversee the application of foresight techniques, ensure continuous iteration, and embed a culture of adaptability throughout the institution. By maintaining a forward-looking posture, the institution can stay ahead of malicious actors, addressing both immediate threats and emerging challenges with precision.

Knowledge Management and Institutional Memory

A learning institution must cultivate a robust institutional memory, supported by effective knowledge management systems. This involves creating a centralised, evolving repository of insights, case studies, and best practices. Tools like AI-driven databases or natural language processing algorithms can help extract insights from previous experiences, allowing future strategies to be informed by past lessons. Moreover, even simple internal communication platforms like Slack or Yammer can play a vital role. These platforms allow staff to ask questions, exchange ideas, and search for relevant insights using keywords, much like Stack Overflow does in programming. Coupled with a good knowledge management system, this allows for rapid problem-solving and the sharing of expertise across the institution, ensuring no critical information is lost.

By maintaining an institutional memory that constantly evolves, the organisation can ensure that the knowledge accumulated over time remains accessible and useful for future generations of staff, which is particularly relevant in a model that relies in large volumes of volunteers.

Leveraging Citizen Science and Open Data Initiatives

To further strengthen its adaptive capacity, the institution could actively engage with open data and citizen science initiatives while addressing one of the most pressing challenges in this field: researchers' access to platform data. By making anonymised data available to the public, researchers, and journalists, the institution can crowdsource insights and patterns it might otherwise miss. Citizen science platforms enable the broader public to participate in detecting disinformation trends, transforming passive audiences into active contributors. This democratises the process and enhances the institution's ability to identify emerging threats through a wider lens.

Access to social media platform data has been repeatedly identified as a critical barrier in addressing synthetic content and disinformation, as platforms often possess unique insights into the mechanics of malicious campaigns. As a recent paper from The Alan Turing Institute emphasises[58], facilitating researcher access to anonymised platform data is essential for developing effective countermeasures. The institution could play a pivotal role in piloting collaborative frameworks that bring together platforms, researchers, and policymakers. Such frameworks would need to prioritise transparency, privacy, and accountability, ensuring data sharing serves the public interest without compromising ethical standards. However, by addressing this systemic issue, the institution could unlock a much better understanding of disinformation dynamics while setting an international benchmark for responsible data sharing.

ABOUT THE AUTHORS

ALEŠ ČÁP

Aleš is a PhD researcher at University College London (UCL). His research focuses on harms caused by Generative AI to democratic processes, and application of collective intelligence to mitigate those harms. By employing a blend of Futures, experimental, and computational methods, his research anticipates potential threats and develops and tests appropriate mitigation strategies.

Before his PhD, Aleš worked as a management consultant with a focus on organisational and institutional design. Projects included supporting the NHS on its transition to Integrated Care System (ICS); merger of large organisations in the UK energy sector; or designing and implementing a new operating model for a UK Ministerial Department. Aleš holds a Master of Science degree in Psychology (Distinction) from UCL.

SIR GEOFF MULGAN

Geoff is a Professor at University College London. He has had a career spanning senior roles in governments, NGOs, foundations and business. He has been directly involved in setting up many organisations in the public sector and civil society, and has experience overseeing venture capital funds and impact investment. He is the author of various books and other writings that included reflections on institutional design, including 'The Art of Public Strategy', 'Big Mind', 'When Science Meets Power', and shorter pieces proposing designs for institutions ranging from data trusts to new global governance entities. He is co-editor in chief of the journal Collective Intelligence and a board member at the Centre for European Policy Studies.

ACKNOWLEDGMENTS

We are deeply grateful to the community of experts who reviewed and provided feedback on an earlier draft of this paper. Their insights have been invaluable in shaping our thinking and strengthening this work. In particular, we thank Marietje Schaake, Sam Gregory, Sam Stockwell, Henry Parker, Alex Fischer, Lisa Witter, Subhajit Basu, Jay Weatherill, Alexander Evans, and Matt Day for their thoughtful contributions.

That said, any errors or omissions are solely the responsibility of the authors.

We hope this paper serves as a useful starting point for further discussion.

APPENDIX A: HARNESSING TENSIONS FOR COLLABORATION

The more granular aspects of internal anatomy could also be understood as "ways of working," or more simply as *how we do things around here*. This encompasses the daily practices, behaviours, and processes that define the operational culture of an organisation. These methods are not just procedural; they are the embodiment of the institution's ethos and task. By intentionally designing and continuously refining these processes, an institution ensures that its core values and purpose are consistently reflected in its actions. Thoughtful design of the granular internal anatomy is an opportunity to translate the institution's ethos into tangible practices.

The institution's design needs to mirror the key values like transparency, promptness and -importantly- collaboration. The anatomy will therefore have to enable it to *exist as a hub*, facilitating collaboration amongst various stakeholders, and balancing the promptness of decision-making with caution and accountability.

Much of the internal anatomy will therefore need to be designed to *harness the tensions* found in the organisation and among its stakeholders into a constructive output without sacrificing its decisiveness and speed of decision-making. In multistakeholder environments, tensions and disagreements are not only inevitable – they are essential. These tensions can drive innovation, problem solving and robust solutions when harnessed effectively.

Similarly, the necessity to balance caution and responsible decision-making means that careful thought needs to be given to the mechanisms that shape and surround the decision-making processes, understanding when to accelerate and when to slow down the process, when to empower people and when to seek debate and challenge.

The following examples illustrate some options of tools, processes, and principles that when implemented, can help harness these tensions productively:

— *ICANN's "Rough Consensus":*

  Rough Consensus, as illustrated in the ICANN example above, is a decision-making model that prioritises open discussion and broad agreement over formal voting procedures. It is characterised by its emphasis on general agreement and the resolution of significant objections, rather than seeking unanimity or strict majority rule. This approach promotes inclusivity and agility, facilitating swift decision-making in dynamic and collaborative environments. By valuing the collective judgement of the group, Rough Consensus encourages wide participation and fosters a sense of shared ownership among stakeholders.

  However, the lack of formal metrics can introduce ambiguity, as the threshold for what constitutes consensus can vary. Additionally, more vocal, senior, or high-status participants may dominate discussions, potentially marginalising some voices and skewing the perceived agreement. Despite these challenges,

Rough Consensus remains an effective mechanism for environments where flexibility and responsiveness are crucial, provided that mechanisms are in place to ensure balanced representation and clarity in the decision-making process.

— *Bridgewater's "Idea Meritocracy":*

Bridgewater Associates, led by Ray Dalio, employs a system known as *radical transparency* and *idea meritocracy*. This approach involves recording all meetings and encouraging employees to openly challenge each other's ideas. Decision-making is based on the best ideas as determined by data and rigorous debate, not hierarchy. This method leverages tension to surface the best solutions, ensuring that the decision-making process is thorough and inclusive. It with its emphasis on recording meetings and encouraging debate, this method also fosters a culture of transparency and accountability.

— *"Disagree and Commit":*

Principle of "disagree and commit" allows team members to voice their concerns but requires everyone to commit to the decision once it is made. This method helps avoid the consensus trap, where lack of consensus leads to inaction. It ensures that diverse opinions are heard, but decision-making is not paralysed by the need for unanimity. Several organisations ranging from Amazon, Netflix, to Intel have been known to utilise this method.

— *"Advice Process":*

The "advice process" empowers individuals to make decisions after consulting all affected parties and those with relevant expertise. This method balances inclusivity with efficiency, ensuring informed decision-making without falling into the consensus trap. Once advice is sought, the decision-maker is not obligated to follow it, allowing for swift, decisive action. Crucially, the advice process allows any individual within the organisation to make decisions, provided they consult with those impacted and knowledgeable in the subject matter.

Buurtzorg, a Dutch home healthcare organisation, exemplifies this approach. Nurses in self-managed teams make decisions by seeking advice from colleagues, ensuring diverse perspectives are considered. This decentralised model has led to high employee satisfaction and improved patient outcomes, demonstrating the advice process's effectiveness in fostering operational efficiency and high-quality care.

— *Bain's "RAPID Framework"*

An example of a more hierarchical option is The RAPID framework by Bain & Company. It's is an effective decision-making mechanism that balances caution with speed. This framework assigns clear roles: recommending (R) actions based on thorough analysis, obtaining agreement (A) from key stakeholders, performing (P) the tasks, incorporating expert input (I), and having a designated decision-maker (D) to ensure accountability.

This structure enhances speed by reducing bottlenecks and promoting clear communication, while maintaining caution through multiple review layers. The RAPID framework's balance of efficiency and thorough vetting makes it suitable for contexts where quick yet well-considered decisions are vital.

APPENDIX B: FOSTERING A CULTURE OF COLLABORATION

To ensure collaboration, accountability, and prompt decision-making, the institution must empower individuals to make informed decisions confidently. This requires robust support systems, targeted training, and a culture that values responsible autonomy to avoid unnecessary delays or backlogs. Achieving this balance demands intentional structuring of internal processes, including clear policies, effective training programs, meaningful rituals, and well-designed reward systems. These elements should be crafted to foster a culture that balances speed with responsibility, ensuring the institution operates efficiently while maintaining its core values.

— *Data and Resource Sharing:*

Ensuring seamless data and resource sharing among stakeholders is fundamental. Singapore's healthcare system exemplifies this with its integrated care approach, often cited as a model of excellence. By sharing real-time patient data among hospitals, clinics, and providers, Singapore ensures coordinated, efficient care, reducing service duplication and improving outcomes.

Similarly, the Global Health Security Agenda (GHSA), with 71 member countries and 12 organisations, demonstrates multilateral collaboration and information sharing. This effort enhances global health security by leveraging shared data and resources among diverse stakeholders. These examples show that with clear objectives and robust frameworks, effective data sharing is achievable for any institution.

— *Collaboration Tools:*

Using digital platforms for communication and project management is crucial. Tools like Slack or Microsoft Teams provide channels for different focus areas, ensuring continuous dialogue. Platforms like Miro or Mural facilitate virtual brainstorming and problem-solving sessions, making it easier for stakeholders to collaborate remotely. Collaborative methods for document creation, like Google Docs or Microsoft Office 365, streamline the process, avoiding the convoluted version management often seen in traditional offline document sharing. Author's experience in organisational transformation has revealed that even renowned organisations often fail to utilise these simple yet highly effective collaboration tools.

— *Conflict Resolution and Feedback:*

To foster a collaborative culture, individuals must be equipped with two key skills: conflict resolution and feedback mechanisms. Effective *conflict resolution* is crucial in multistakeholder environments where diverse perspectives can lead to disagreements. Buurtzorg, the Dutch home healthcare organisation,

exemplifies this by training its self-managed teams in structured dialogue and mutual respect, ensuring disputes are resolved promptly and do not hinder performance or care quality.

— Equally important are robust *feedback mechanisms*, where team members regularly provide constructive feedback. This practice fosters continuous improvement and accountability, as seen in high-performing teams across various sectors. Google's research identified psychological safety—the belief that one will not be punished for speaking up—as the number one predictor of high-performing teams. Regular feedback and effective conflict resolution are essential for building this psychological safety.

These skills are learnable. Institutions must provide training and support, set clear expectations, and establish mechanisms that reinforce these practices. By embedding conflict resolution and feedback mechanisms into the organisational culture, teams can maintain high standards of performance and adapt swiftly to changing conditions, ensuring a cohesive, responsive, and high-performing organisational culture.

— *Leading by Example:*

Leadership must exemplify the behaviours they advocate to foster a genuinely collaborative and high-performing organisational culture. When leaders consistently demonstrate the values and practices they promote, such as transparency, accountability, and impartiality, they set a powerful precedent for the entire organisation. Conversely, any leader failing to embody these principles can significantly undermine the institution's credibility and cohesion, leading to a detrimental impact on morale and performance.

— *Reward Mechanism:*

Another good example of promoting key behaviours through deliberate design decisions is Google's *Courageous Penguin Award*. Inspired by penguins standing at the edge of an iceberg and contemplating whether to jump into the water below, the first penguin to leap showcases bravery by taking a risk without knowing the outcome. Similarly, Google recognises employees who demonstrate such courage, encouraging a culture where individuals are willing to take risks and try new things, even without guaranteed success. Institutions that understand and meaningfully reward key desired behaviours are the ones that successfully cultivate the cultures and behaviours they aim to see.

— *Designing a Collaborative Culture:*

Frederic Laloux's book "Reinventing Organisations" provides numerous examples of organisational policies that foster trust and encourage embracing accountability. One notable example is a company where employees can use company cars without needing permission. Most organisations lock their valuable resources away and require multiple levels of permissions, but this 'open resource' approach communicates and fosters a culture of trust and empowerment. Institutional architects must think like designers, creating

environments that encourage desired behaviours rather than simply demanding them.

Once the institution has agreed the final ethos and key behaviours, it can begin to think about its internal anatomy and develop creative policies, mechanisms, and any structural enhancements that mirror its unique values and needs.

APPENDIX C: EMPOWERING PROACTIVE SOLUTIONS: BUFFER DELAYS

Introducing proactive measures into the discussion of power and mandate is essential because key frontline actors such as social media platforms may need to be empowered to act. It's not just about holding these platforms accountable; sometimes, authorities may need to *empower* or even *compel* them to adopt more proactive measures, such as pre-monitoring digital content. This would enable social media companies to act swiftly against harmful disinformation before it spreads, ensuring that the integrity of democratic processes is upheld during critical moments. This idea of shared responsibility underscores how vital it is to integrate social media platforms into the broader framework of electoral integrity, ensuring they have both the power and obligation to act effectively. Given the speed at which content spreads, institutions need proactive, real-time response mechanisms to safeguard against disinformation, especially during high-risk periods like election campaigns.

Southport as a Case for Pre-monitoring?

A recent example worth considering regarding the executive power of is how an electoral integrity body could handle extremely emotional, inflammatory, highly emotional, and potentially hyper-viral disinformation scenarios. An example of such case is the July 2024 stabbing in Southport, UK, where misinformation circulated online about a killer of three children being an illegal immigrant, even fabricating a name for the alleged killer. Preying on public fears and the emotional charge of the situation, this piece of disinformation gained country-wide traction and went viral. It demonstrates just how serious the potential harm is disinformation can cause, exacerbating tensions and undermining public trust, and in this case, leading to nationwide riots and violence. In times when falsehoods can spread rapidly and widely, it may be necessary to explore proactive remedies such as pre-monitoring technologies capable of identifying and assessing harmful content *before* it goes viral.

Ethics of Pre-monitoring:

The ethics and risks of pre-monitoring content for misinformation before publication are undoubtedly debatable. While many authorities such as the European union (EU) have traditionally been opposed to widespread pre-upload monitoring, this momentary shift towards precautionary measures during elections, if implemented and managed cautiously, could be a pragmatic and effective way to safeguard democratic integrity without violating free speech. It is arguably an area that deserves further research - and indeed debate. In the EU example, the Digital Services Act (DSA) explicitly prohibits general monitoring obligations for online platforms. This means that platforms are not required to actively monitor all user-generated content or search for illegal content across their services. However, the Court of Justice of the European Union (CJEU) differentiates between general monitoring obligations and monitoring obligations in specific cases, which may be ordered by national authorities. Recital 28 of the proposal now expressly upholds this distinction. This means that if countries find it appropriate, certain mitigating measures can be applied during electoral periods.

Meta's existing technologies already scan for prohibited content such as child pornography or nudity, demonstrating that the infrastructure for pre-monitoring mechanisms exist. This wouldn't be radically new. Meta has experience using AI and hashes (digital footprints found in images and videos) to collaborate with other internet firms to prevent spread of terrorist content online using an open-source Hasher-Matcher-Actioner tool[59]. In the run-up to elections, particularly during the final few days, there could be room for pre-monitoring of content that is flagged as extremely high-risk or aligned with known disinformation narratives. This method could be employed in a targeted and time-limited manner, ensuring that freedom of expression is preserved but with a slight tilt towards safety during high-stakes periods.

Should Buffer Delays be Utilised?

This approach could involve *buffer delays*—temporary holds on content flagged as likely disinformation with extremely high emotional salience and virality potential, giving human reviewers the time to rapidly assess whether the content is harmful. This would provide another layer of protection, a safety net as well as a deterrent, against the rapid spread of disinformation. Importantly, the system would need to minimise the risk of false positives as well as false negatives and maintain a balance between freedom of expression and electoral integrity. Mistakes are inevitable, but if through progress and development such errors can become marginal, such a system might prove highly effective in mitigating the harms. Here, the present institution's role may be to drive the debate towards defining the thresholds for action—what constitutes sufficient risk to justify delaying or freezing content, and under what circumstances pre-monitoring becomes necessary.

ENDNOTES

1 Mustafa Suleyman. *The coming wave: technology, power, and the twenty-first century's greatest dilemma*. Crown, 2023.

2 Nina Schick. *Deep fakes and the infocalypse: What you urgently need to know*. Hachette UK, 2020.

3 Sam Stockwell. "AI-Enabled Influence Operations: Safeguarding Future Elections". In: *The Alan Turing Institute* (2024).

4 Serena Iacobucci et al. "Deepfakes unmasked: the effects of information priming and bullshit receptivity on deepfake recognition and sharing intention". In: *Cyberpsychology, behavior, and social networking* 24.3 (2021), pp. 194–202.

5 Nina Schick. *Deep fakes and the infocalypse: What you urgently need to know*. Hachette UK, 2020.

6 Sadeghi McKenzie. "Tracking AI-enabled Misinformation: 957 'Unreliable AI-Generated News' Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools". In: (2024).

7 Bobby Chesney and Danielle Citron. "Deep fakes: A looming challenge for privacy, democracy, and national security". In: *Calif. L. Rev.* 107 (2019), p. 1753.

8 Mika Westerlund. "The emergence of deepfake technology: A review". In: *Technology innovation management review* 9.11 (2019).

9 Logically. "Combating Deepfakes Disinformation: Proposed Strategies for the UK's Digital Future". In: *n/a* (2024).

10 Maria Pawelec. "Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions". In: *Digital society* 1.2 (2022), p. 19.

11 Simone Chambers. "Truth, deliberative democracy, and the virtues of accuracy: is fake news destroying the public sphere?" In: *Political Studies* 69.1 (2021), pp. 147–163.

12 Mateusz Labuz and Christopher Nehring. "On the way to deep fake democracy? Deep fakes in election campaigns in 2023". en. In: *European Political Science* (Apr. 2024). issn: 1680-4333, 1682-0983. doi: 10.1057/s41304-024-00482-9. url: https://link.springer.com/10.1057/s41304-024-00482-9 (visited on 05/08/2024).

13 Morgan Meaker. *Slovakia's election deepfakes show ai is a danger to democracy*. 2023.

14 Nina Schick. *Deep fakes and the infocalypse: What you urgently need to know*. Hachette UK, 2020.

15 Sam Stockwell. "AI-Enabled Influence Operations: Safeguarding Future Elections". In: *The Alan Turing Institute* (2024).

16 Meryl Sebastian. "AI and deepfakes blur reality in India elections". In: *BBC* (2024).

17 James A Robinson and Daron Acemoglu. *Why nations fail: The origins of power, prosperity and poverty*. Profile London, 2012.

18 Daron Acemoglu and James A Robinson. *The narrow corridor: How nations struggle for liberty*. Penguin UK, 2019.

19 Maria Pawelec. "Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions". In: *Digital society* 1.2 (2022), p. 19.

20 n/a. *Canada's plan to protect democracy*. 2019. url: https://www.canada.ca/en/democratic-institutions/services/protecting-democracy.html#2.

21 SGDSN. *Service de vigilance et protection contre les ing´erences num´eriques ´etrang`eres*. 2022. url: https://www.sgdsn.gouv.fr/notre-organisation/composantes/servicede-vigilance-et-protection-contre-les-ingerences-numeriques.

22 EU DisinfoLab. *Securing information integrity in the upcoming French presidential election: a catalogue of initiatives*. 2022. url: https://www.disinfo.eu/publications/ securing-information-integrity-in-the-upcoming-french-presidential-election/.

23 EU DisinfoLab. *Securing information integrity in the upcoming French presidential election: a catalogue of initiatives*. 2022. url: https://www.disinfo.eu/publications/ securing-information-integrity-in-the-upcoming-french-presidential-election/.

24 VIGINUM. *MATRYOSHKA A pro-Russian campaign targeting media and the fact-checking community*. 2024. url: https://www.sgdsn.gouv.fr/files/files/20240611_NP_SGDSN_VIGINUM_Matriochka_EN_VF.pdf.

25 Douglass C North. *Institutions, institutional change and economic performance*. Cambridge university press, 1990.

26 Connor Dunlop Merlin Stein. "Safe before sale". In: *n/a* (2023).

27 Marietje Schaake. *The Tech Coup: How to Save Democracy from Silicon Valley*. Princeton University Press, 2024.

28 Harris Gleckman. *Multistakeholder governance and democracy: A global challenge*. Routledge, 2018.

29 The Christchurch Call. *The Christchurch Call Commitments*. 2019. url: https://www.christchurchcall.org/the-christchurch-call-commitments/.

30 Philip Schleifer. "Varieties of multistakeholder governance: selecting legitimation strategies in transnational sustainability politics". In: *Globalizations* 16.1 (2019), pp. 50–66.

31 Nicola Palladino and Mauro Santaniello. "Foundations, Pitfalls, and Assessment of Multistakeholder Governance". In: *Legitimacy, Power, and Inequalities in the Multistakeholder Internet Governance: Analyzing IANA Transition*. Cham: Springer International Publishing, 2021, pp. 21–42. isbn: 978-3-030-56131-4. doi: 10.1007/978- 3- 030- 56131- 4_2. url: https://doi.org/10.1007/978-3-030-56131-4_2.

32 Luke Harding Dan Sabbagh and Andrew Roth. *Russia report reveals UK government failed to investigate Kremlin interference*. 2020. url: https://www.theguardian.com/world/2020/jul/21/russia-report-reveals-uk-government-failed-to-addresskremlin-interference-scottish-referendum-brexit.

33 Nina Schick. *Deep fakes and the infocalypse: What you urgently need to know*. Hachette UK, 2020.

34 Cameron Martel et al. "Crowds can effectively identify misinformation at scale". In: *Perspectives on Psychological Science* 19.2 (2024), pp. 477–488.

35 Yvonne McDermott-Rees. "Trust in User-Generated Evidence in an Era of Deepfakes: Insights from the TRUE Project". In: *Conference paper, Deepfakes Law, City University*, 2024 (2024).

36 et al. Martel Cameron. "Harnessing Partisan Motives to Solve the Misinformation Problem". In: *Conference paper, ACM Collective Intelligence*, Boston, 2024 (2024).

37 Thomas W Malone. Superminds: *How hyperconnectivity is changing the way we solve problems*. Simon and Schuster, 2018.

38 Nina Schick. *Deep fakes and the infocalypse: What you urgently need to know*. Hachette UK, 2020.

39 Md Shohel Rana et al. "Deepfake detection: A systematic literature review". In: *IEEE access* 10 (2022), pp. 25494–25513.

40 Matthew Groh et al. "Deepfake detection by human crowds, machines, and machineinformed crowds". en. In: *Proceedings of the National Academy of Sciences* 119.1 (Jan. 2022), e2110013119. issn: 0027-8424, 1091-6490. doi: 10.1073/pnas.2110013119. url: https://pnas.org/doi/full/10.1073/pnas.2110013119 (visited on 10/20/2023).

41 Nakhoon Choi and Heeyoul Kim. "DDS: deepfake detection system through collective intelligence and deep-learning model in blockchain environment". In: *Applied Sciences* 13.4 (2023), p. 2122.

42 Logically. "Combating Deepfakes Disinformation: Proposed Strategies for the UK's Digital Future". In: *n/a* (2024).

43 Logically. "Combating Deepfakes Disinformation: Proposed Strategies for the UK's Digital Future". In: *n/a* (2024).

44 E. Glen Weyl, Audrey Tang, and the Plurality Community. *Plurality: The Future of Collaborative Technology and Democracy*. 2023. url: https://github.com/pluralitybook/plurality/blob/main/contents/english.

45 Kim Sengupta. "Meet the Elves, Lithuania's digital citizen army confronting Russian trolls". In: *n/a* (2019).

46 Nathan Walter et al. "Fact-Checking: A Meta-Analysis of What Works and for Whom". en. In: *Political Communication* 37.3 (May 2020), pp. 350–375. issn: 1058-4609, 1091-7675. doi: 10.1080/10584609.2019.1668894. url: https://www.tandfonline.com/doi/full/10.1080/10584609.2019.1668894 (visited on 02/06/2024).

47 Jana Siebert and Johannes Ulrich Siebert. "Effective mitigation of the belief perseverance bias after the retraction of misinformation: Awareness training and counter-speech". In: *Plos one* 18.3 (2023), e0282202.

48 Logically. "Combating Deepfakes Disinformation: Proposed Strategies for the UK's Digital Future". In: *n/a* (2024).

49 Joseph Bambridge. *British MPs fear we can't stop election deepfakes. They're right*. 2024. url: https://www.politico.eu/article/united-kingdom-deepfakes-electionrishi-sunak-keir-starmer-sadiq-khan/.

50 n/a. *Battle With X Over Starmer Deepfake Highlights UK Election Worry*. 2024. url: https://www.bloomberg.com/news/articles/2024-03-15/battle-with-x-overstarmer-deepfake-highlights-uk-election-worry?embedded-checkout=true.

51 n/a. *VIGINUM: French defense against cyber-attacks fake news*. 2024. url: https://www.aapafrance.org/viginum-french-defense-against-cyber-attacks-fake-news/.

52 ITV. *GUK riots: Jail sentences and community orders handed to individuals involved in unrest*. 2024. url: https://www.itv.com/news/2024-08-07/uk-riots-jail-sentences-handed-to-individuals-involved-in-unrest.

53 E. Glen Weyl, Audrey Tang, and the Plurality Community. *Plurality: The Future of Collaborative Technology and Democracy*. 2023. url: https://github.com/pluralitybook/plurality/blob/main/contents/english.

54 Geoff Mulgan. "Organisational architecture: Ideas for an emergent discipline". In: n/a (2022).

55 DOJ. *Justice Department Disrupts Covert Russian Government-Sponsored Foreign Malign Influence Operation Targeting Audiences in the United States and Elsewhere*. 2024. url: https://www.justice.gov/opa/pr/justice-department-disrupts-covertrussian-government-sponsored-foreign-malign-influence.

56 Higgins Tamlin. *The EU AI Act: concerns and criticism*. 2023. url: https://www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2023/04/the-eu-ai-act--concerns-and-criticism.html.

57 Martens Bertin. *The European Union AI Act: premature or precocious regulation?* 2023. url: https://www.bruegel.org/analysis/european-union-ai-act-premature-orprecocious-regulation#:~:text=The%20incompleteness%20of%20the%20AI, cit. %20..

58 Sam Stockwell. "AI-Enabled Influence Operations: Safeguarding Future Elections". In: *The Alan Turing Institute* (2024).

59 n/a. *Meta shares Hasher-Matcher-Actioner tool for detecting and deleting terrorist content online*. 2022. url: https://siliconangle.com/2022/12/13/meta-shares-hashermatcher-actioner-tool-detecting-deleting-terrorist-content-online/.

THE INSTITUTIONAL ARCHITECTURE LAB

WWW.TIAL.ORG