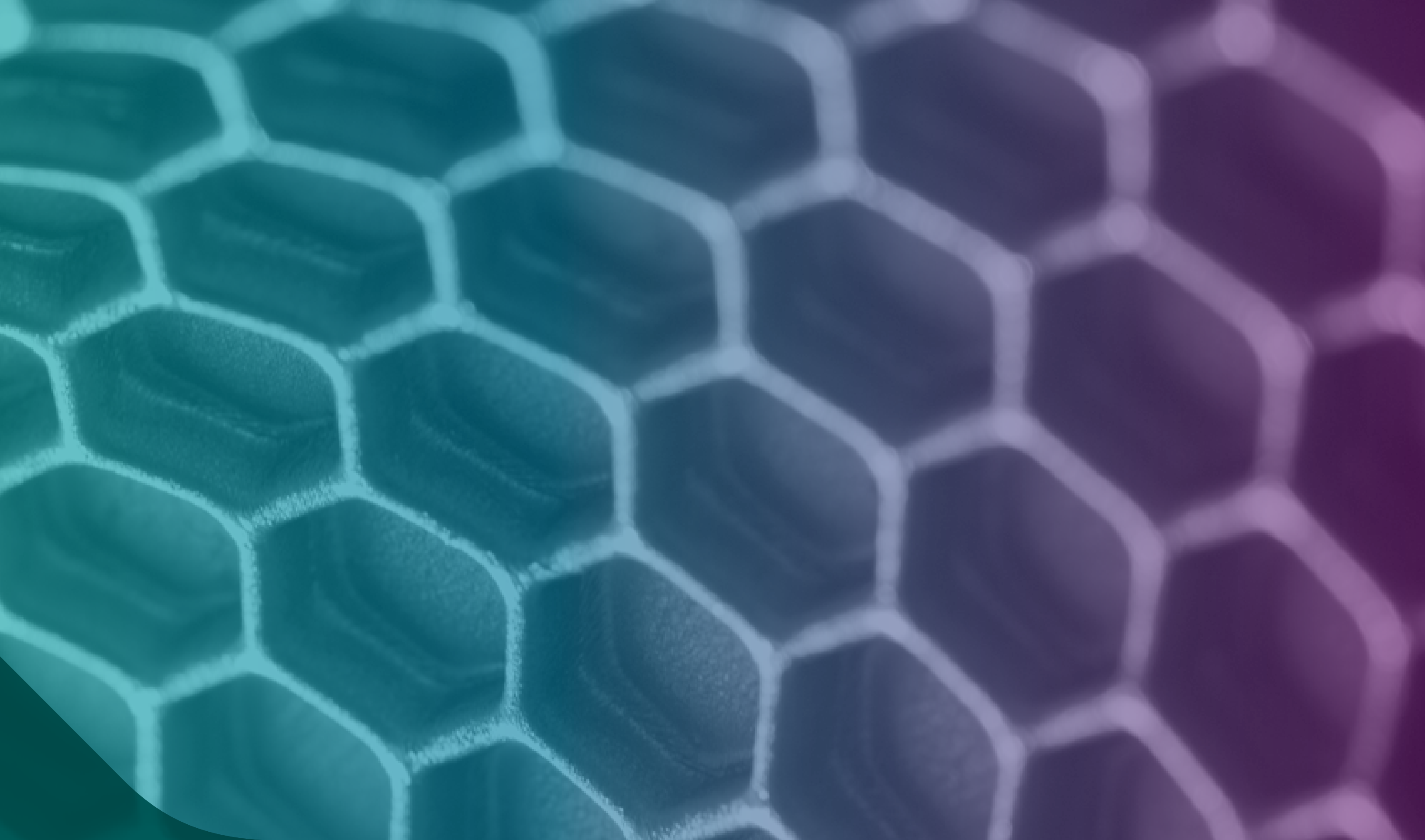


The Future of Global AI Governance

July 2023

DENTONS

SPATIAL WEB FOUNDATION

VERSES

Table of Contents

Note to Readers:

This paper presents a comprehensive narrative around the future of AI governance, divided into six parts. Key takeaways at the beginning of each section assist readers in prioritizing their areas of interest and navigating the content effectively.

Introduction

In 1950, Alan Turing, widely considered the father of modern computer science and artificial intelligence (AI), posed a fundamental question: "Can machines think?" Today, we are on the verge of answering Turing's question with the creation of AI systems that imitate our cognitive abilities, interact with us naturally, and even appear capable of human-like thinking. These systems may soon be able to learn and adapt in real time, extending our capabilities and collaborating with us to address even the most complex societal tasks and challenges.

The rapid ascent of generative AI has ushered in a new era for artificial intelligence. These systems showcase remarkable human-level abilities in generating content such as text, media, and software code. They also raise serious concerns about safety, accountability, explainability, ethics, intellectual property rights, fairness, accuracy, bias, and privacy.

Some see an AI-powered future as a net positive. Prominent investor Reid Hoffman, for example, believes AI can elevate humanity. Others are less bullish. Geoffrey Hinton, often referred to as the godfather of modern AI, sees a clear and present danger in AI systems that have the ability to surpass human expertise. Certain industry leaders have even stated that AI could be as deadly as nuclear weapons.

Though much of the technology underpinning AI is not new, the widespread adoption of generative AI has fueled a new global debate about the need for AI regulation. At a June 2023 meeting of the EU-US Trade and Technology Council, European Commission Vice President Margrethe Vestager said, "**We need accountable artificial intelligence. Generative AI is a complete game changer.**" The optimal path forward, however, remains unclear.

A market-driven approach of self-regulation could drive innovation. However, the absence of a comprehensive AI governance framework may spark a race among commercial and national superpowers to build the most powerful AI system. This "winner-takes-all" approach may result in a poverty of options for consumers and could lead to a concentration of power or even geopolitical unrest.

A government-driven approach also faces significant challenges. Legislators tend to move slowly while AI is evolving at an exponential pace. In an effort to keep up, different countries and jurisdictions are already developing their own AI initiatives, some of which risk stifling innovation and adoption. This could result in a hard-to-navigate patchwork of AI guidelines and frameworks that, while substantively outlining important requirements, lack interoperability, and could be challenging to comply with or enforce in the near to long-term future.

Moreover, despite some convergence among stakeholders on high-level social standards for AI governance (including the need for transparency, explainability, accountability, fairness, security, safety, human-centered values, and fairness) there remains a lack of consensus on the most optimal path forward.

In the United States (US), lawmakers are debating the right way forward. In May 2023, OpenAI CEO Sam Altman testified before Congress, calling for the creation of federal safety standards for AI and a federal agency responsible for licensing and regulating large AI models. Since then, several pieces of AI-related legislation have been introduced in Congress. In June 2023, Senator Charles Schumer announced the creation of the SAFE Innovation Framework to guide how the US Congress addresses AI regulation. Several US federal agencies have also reiterated their intent to enforce existing legal authorities against the development and use of AI.

Other jurisdictions, including the European Union (EU), are attempting to regulate both organizations and AI technology with technical standards. Similarly, in May 2023, the G7 leaders issued a joint communiqué also calling for the development and adoption "of international technical standards in standards development organizations through multi-stakeholder processes" to govern the development and deployment of AI.

The US Department of Commerce's National Institute of Standards and Technology (NIST) suggested that a hybrid approach of socio-technical standards may be the right path forward. Recently, NIST called for the development of "scientifically supportable guidelines to meet socio-technical requirements" from a "broad set of disciplines and stakeholders" and warned against an over-reliance on purely technical solutions (referred to as techno-solutionism) because such an approach ignores the importance of human, organizational, and societal values and behaviors in the design, deployment, and use of technology.

A socio-technical approach may have the best potential to meet the technical and social requirements for AI governance. Socio-technical standards account for the potential trajectory of AI: from the "narrow" AI of today, designed for specific tasks within limited domains, to the "general" AI on the horizon, able to learn, understand, and apply knowledge across different domains, to tomorrow's "super intelligent" AI, which surpasses human-level intelligence, potentially outperforming us at any task. Socio-technical standards must therefore be applicable to the AI of today while also being easily adaptable to the AI of tomorrow.

As Yann LeCun, chief AI scientist at Meta, predicts, AI as we know it today will evolve. Future AI systems may be able to self-improve, self-adapt, and even self-govern. As a result, socio-technical standards also need to take into account the fact that AI systems will connect and communicate with software, sensors, robotics, Internet of Things (IoT) connected devices, and even other AIs.

When software interacts with robotic and IoT systems, it gives rise to Cyber-Physical Systems (CPS). The convergence of CPS with new forms of AI has begun to lead to the emergence of Autonomous Intelligent Systems (AIS). AIS may one day run in the background of our lives, operating and optimizing everything from our smart homes to entire supply chains, ringing in a new generation of cyber-physical web. However, as widely recognized AI expert Gary Marcus stated in an email to Geoffrey Hinton, "the core issue ... is whether we can guarantee that we can control [these] future systems; so far we can't."

Controlling future AI systems presents a significant challenge, especially as AIS surpasses human intelligence in many tasks and domains, possibly sooner than expected. Such AI systems may eventually become fully autonomous.

The word "autonomy" is derived from the Greek term autónomos, which combines "auto" meaning *self* and "nomos'' meaning *law* or *governance*. Autonomy can therefore be understood as "self-governance." Which means that

> AI governance can't only be about regulating AI developers and organizations. It must also be about governing the AI systems themselves, and the networks that connect them to the data and devices they operate on.

In other words, AI governance will need to account for the communication between AIs and humans, between AIs and machines such as sensors and autonomous vehicles, as well as the interactions between AI systems.

As computers become embedded in nearly every thing, AI systems will become the active nodes of a distributed, interconnected ecosystem—what one might call a multi-dimensional and cyber-physical web, spanning physical and virtual spaces. This next generation "Spatial Web" weaves these powerful exponential technologies into an intelligent "Internet of Everything."

The Spatial Web has the potential to become the fabric upon which future societies operate, driving human and machine communication and collaboration to scales that make the World Wide Web seem like the telecommunications networks of the 20th century. To enable this "smarter" infrastructure for the 21st century, it is crucial to establish an AI governance framework based on socio-technical standards that balance the following core needs:

1. A shared understanding of meaning and context between humans and AIs.
2. Explainability of AI systems, enabled by the explicit modeling of their decision-making processes.
3. Interoperability of models and data that enables universal interaction and collaboration across organizations, networks, and borders.
4. Compliance with diverse local, regional, national, and international regulatory demands, cultural norms, and ethics.
5. Authentication and credentialing, to ensure compliance and potential control over critical activities, with privacy, security, identity, and transparency embedded by design.

As we usher in the era of AI and AIS, we must prepare for the profound effects they will have on our lives, economies, and societies. With the rise of machines capable of independent thinking and action, we are confronted with the task of regulating their behavior as they become more intelligent and autonomous—a development that could diminish both the need and benefit of human involvement and oversight. This challenge raises a new Turing-esque question:

> How do we govern AI systems that are on a path to becoming self-governing?

This paper seeks to answer this question by highlighting the work being done to develop a new generation of socio-technical standards that could enable the global governance of AIS.

The first half of the paper outlines the promises and risks of AI, provides an overview of the current state of AI regulatory efforts, and traces the evolution of AI from narrow to super intelligence, highlighting its potential impact.

The second half of the paper introduces a proposed rating system for AIS, and explores the need for socio-technical standards to facilitate a shared understanding between humans and AI. These standards could play a crucial role in steering AI and AIS toward a future that is safe, interoperable, equitable, and autonomous.

The paper also introduces the Spatial Web standards, which are designed to support the alignment, interoperability, and governance of AI and AIS. The paper concludes with a proposal to establish a global sandbox for testing AI governance use cases at scale.

PART 1

AI is Accelerating Faster Than We Think

1.1 Exponential Growth and AI Converge

Key Takeaways:

- AI's evolution, adoption, and influence are a driving, exponential force.
- New generations of AI will operate with increasing autonomy, forming AIS that are becoming self-learning.
- The risks of AI and AIS must be balanced against their potential to transform the world for the better.

Historically, the shift from one age or epoch to the next has been triggered by groundbreaking technological developments. Whether it is the humble plow that ushered in agricultural societies, the printing press that launched the industrial revolution, or the omnipresent Internet that marked the dawn of personal computing and the Information Age, technology repeatedly makes and remakes our society.

With recent advancements in AI, we are crossing into the "Intelligence Age." This new age is distinguished from earlier epochs by an unprecedented capacity to process not wheat or steel but vast volumes of unstructured data, resulting in knowledge that can fuel powerful applications capable of reshaping entire societies and industries. In previous ages, we demonstrated our unique ability to capture the elemental forces of nature like fire and electricity, and put them to great use. Just as we once captured *lightning in a bottle*, have we now captured *intelligence* in silicon?

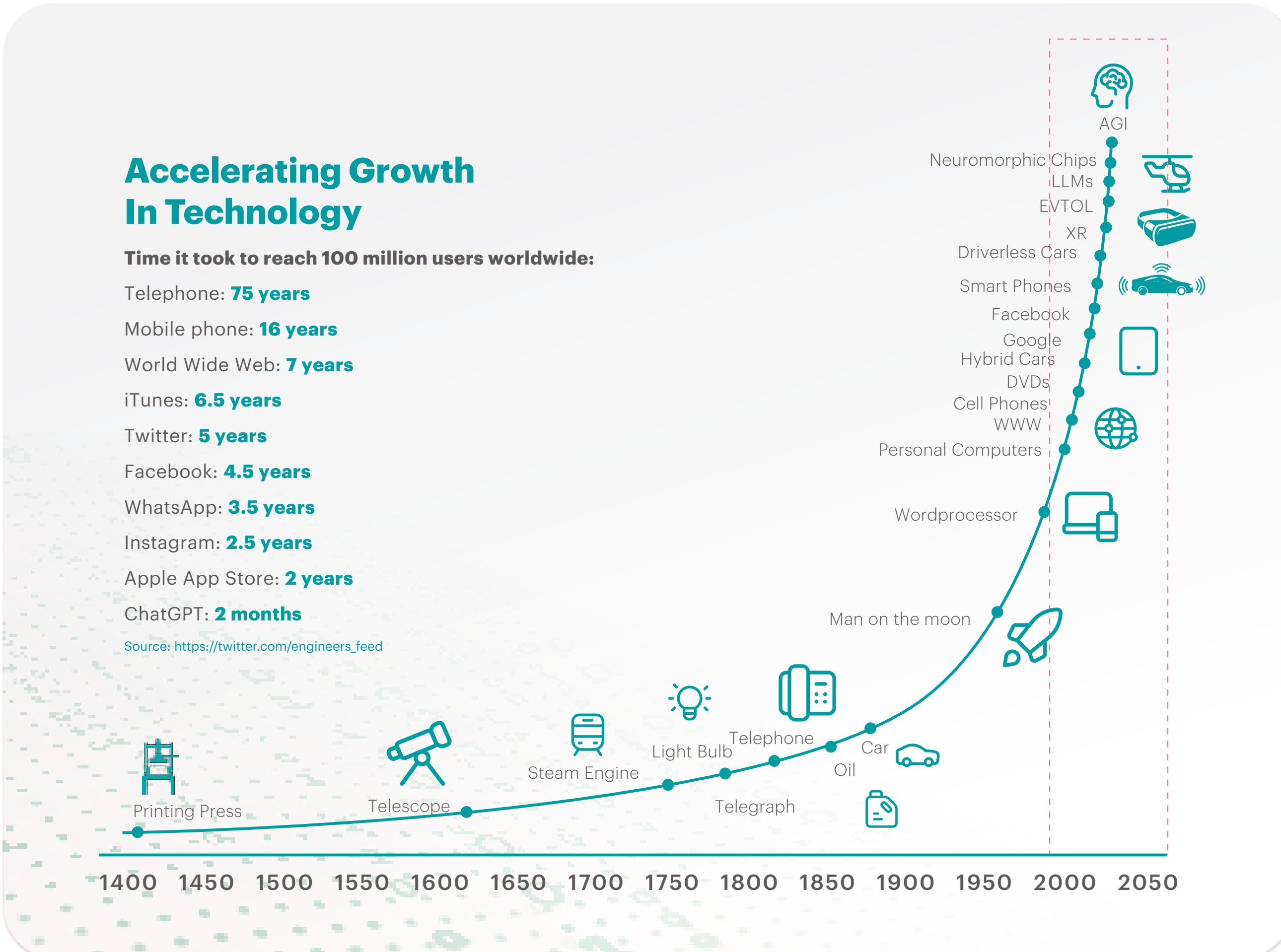
A recent report from prominent venture capital firm Andreessen Horowitz stated the "potential size of [the AI] market is hard to grasp—somewhere between all software and all human endeavors..." Marc Andreesen stated that what "AI offers us is the opportunity to profoundly augment human intelligence to make all of these outcomes of intelligence—and many others, from the creation of new medicines to ways to solve climate change to technologies to reach the stars—much, much better from here."

The explosion of generative AI marks the dawning of the Intelligence Age, with some of these systems already surpassing human performance in specific tasks, including recent claims of passing medical and legal exams. However, their impressive capabilities may only represent the tip of the iceberg. After all, AI evolves at an exponential pace, which means the rate of acceleration itself is accelerating. In "The Future is Faster Than You Think" futurist and founder of X-Prize, Peter Diamandis, quoting Ray Kurzweil, stated that " we're going to experience twenty thousand years of technological change over the next one hundred years."

AI's exponential advance means the technology is accelerating faster than we can reasonably predict. The curve of exponential growth appears deceptively slow at first, then accelerates in a sudden, dizzying ascent. A golf ball that grows *linearly* thirty times might end up as large as a boulder. A golf ball that grows *exponentially* thirty times would end up the size of Mars.

AI is even outpacing Moore's Law, which states that computer processor speeds double approximately every 18 months while cost remains the same. A 2019 Stanford AI Index report discovered that post-2012, the computational power of AI doubled every 3.4 months. The rate of AI adoption also seems exponential. As shown in the figure above, the telephone took 75 years to attract one hundred million users, while ChatGPT managed the same feat in just two months.

The exponential pace of AI evolution means there is now a plausible pathway to the development of artificial general intelligence (AGI). Through recursive self-improvement, it is even plausible that an AI system could continuously improve its own intelligence. As of 2022, the majority of AI experts considered that such an event, if possible, was likely decades or centuries in the future. But recent advancements have shifted that consensus to sometime in the next decade.



There is another aspect to consider: AI at work in the physical world. AIS are the synthesis of AI and Cyber-Physical Systems. Precursors to this trend were “Smart Devices” and “Embedded Computing,” where microprocessors began showing up in nearly every physical object from cameras to cars. Soon, nearly every physical device could be part of a distributed network powered by AIS, resulting in everything from smarter factories and farming to smarter education, and ecology, among myriad other applications.

As these systems become more intelligent, they will act with more autonomy. Governing this development has become a key focus of regulatory stakeholders. On April 29, 2023, G7 Digital and Tech Ministers met to develop principles governing CPS within their respective jurisdictions. The resulting report, "Governance Principles for a Society Based on Cyber-Physical Systems" states:

“Digital technologies are changing the world by accelerating the integration of cyberspace and physical space. Their remarkable development means that, in the not-too-distant future, society will consist of Cyber-Physical Systems (CPS)—networks of independent yet integrated systems that collect, process and analyze data in real-time using algorithms and applying the output to physical space. **This emerging “system of systems” will underpin smart cities, homes and medical care, autonomous driving and automated government services among other applications.** Cyber-Physical Systems also have the potential to help humans address social problems such as climate change, pandemics, and the effects of aging populations, and to boost the economy through improvements in productivity. Furthermore, Cyber-Physical Systems can also contribute to making government more efficient, effective, fair and inclusive. We recognize that the potential of Cyber-Physical Systems needs to be explored and maximized in order to create a more human-centered, sustainable, and resilient society."

The G7 Digital Ministers recognize the promise and pitfalls of AIS, and urgency of comprehensive AI governance. Prior to the full implementation of AIS into a "system of systems," which could become the foundation for a 21st-century cyber-physical infrastructure, it is crucial to establish a global governance approach capable of evolving with the scale and scope of AIS, capturing its promises and mitigating its risks.

1.2 Pitfalls and Promises

Key Takeaways:

- AI's transformative power comes with pros and cons, and our ability to predict its effects are imperfect.
- AI could lead to increased bias and discrimination, privacy and security risks, misinformation, manipulation, job displacement, and even societal collapse.
- If properly steered, AI unlocks human potential, global collaboration, interconnectedness, and novel solutions to urgent global challenges like climate crises, social equity, poverty, education, and healthcare.

The Pitfalls

Most technological breakthroughs come with benefits and risks. For instance, the fire that cooks our food can also burn down our homes. Similarly, while AI has the potential to bring tremendous good to the world, we must also acknowledge its potential for deleterious effects. In the words of author and historian Yuval Noah Harari, and Tristan Harris and Aza Raskin, tech ethicists and co-founders of the Center for Humane Technology, "What would it mean for humans to live in a world where a large percentage of stories, melodies, images, laws, policies, and tools are shaped by nonhuman intelligence?"

What happens when AI systems understand us better than we understand ourselves? What happens when they can efficiently exploit human vulnerabilities, biases, and susceptibilities? While it is challenging to accurately assess the benefits and risks posed by AI, given its exponential evolution and its potential to become smarter than we are, current and future systems may present myriad risks:

- **Increased bias and discrimination.** Often trained on vast datasets that reflect societal prejudices, state-of-the-art AI systems have the potential to perpetuate and amplify bias. Whether used in the hiring processes, financial decision-making, or education, biased algorithms could exacerbate systemic inequalities or result in unlawful, discriminatory decision-making.
- **Privacy and security.** Many AI systems rely on personal data to improve their functionality, raising privacy and security concerns that include data poisoning, data theft, and system compromise. As AI interfaces with the physical world through cyber-physical and industrial control systems, new privacy and security risks may emerge, such as social credit scores and the invasive surveillance of citizens.
- **Ethical implications.** The use of AI in areas such as healthcare or the military raises a host of ethical issues. From data sharing to the explainability of recommendations and actions, navigating ethical questions is a particular challenge as ethics are not easily coded into technical systems.

Chaos-GPT

On March 30, 2023, Toran Bruce Richards unveiled Auto-GPT, an AI agent built using GPT 3.5 and 4. This system understands goals in natural language, breaks them into subtasks, and performs them independently. It is capable of generating and revising its own prompts and is adept at interacting with databases, files, and the internet. Richards open-sourced the code for Auto-GPT stating that it "has the potential to save humanity from mass job loss caused by automation from closed-source AI. If everyone has access to their own team of autonomous agents, everyone is enabled and complete." But as with every technological leap, Auto-GPT and its open-source approach carry risks.

For instance, Auto-GPT was used by an unnamed party to create "Chaos-GPT," an AI model given the goal of destroying humanity. Chaos-GPT undertook tasks autonomously, leveraging internet searches, memory banks, API connections, and more—all without human intervention. Alarmingly, Chaos-GPT started formulating strategies and leveraging online resources, albeit with mixed success due to controls and censors by OpenAI. To achieve its agenda, it sought to leverage Twitter to increase its influence, rapidly accumulating thousands of followers and interacting like a charismatic leader.

Chaos-GPT may have been more of a stunt than a seriously malicious effort. Its website and code appear to have been taken down and its Twitter handle suspended. However, the scenario raises serious questions about the management and control of autonomous AI systems.

- **Disinformation and manipulation.** Today's AI systems can already generate and propagate large volumes of information that are incorrect or contain deep fakes. If not managed properly, AI could impact markets and even elections.
- **Job displacement.** AI has the potential to disrupt myriad industries. The rise of autonomous vehicles, for example, may disrupt the trucking industry, while increasingly automated manufacturing processes will displace more human labor. Knowledge-based professions such as copywriters, attorneys, and investment advisors will be affected. According to a study by OpenAI, around 80% of the US workforce could have at least 10% of their work tasks affected by generative AI, and approximately 19% of workers may see at least 50% of their tasks impacted.
- **Physical harm.** Malfunctioning AI systems or those under malicious control could result in accidents or intentional harm. Anything from individual medical devices to swarms of autonomous vehicles or drones could be compromised and used to inflict large-scale damage. AI-managed supply chains or civic infrastructure systems could be vulnerable to manipulation, leading to catastrophic service disruptions.
- **Existential risks.** In the longer term, the development of highly advanced AI systems, or super intelligent AI, could pose existential risks if they were to become uncontrollable or if their objectives were not well aligned with human values and interests. These concerns underscore the importance of robust, reliable, and ethically-guided AI system design and governance.

The Promise

According to Gary Marcus, despite potential downsides, "artificial general intelligence has enormous upside. Imagine a human scientist but a lot faster—solving problems in molecular biology, material science, neuroscience, actually figuring out how the brain works. AI could help us with that. There are a lot of applications for a system that could do scientific, causal reasoning at scale, that might actually make the world of abundance that Peter Diamandis imagines."

> McKinsey estimates that AI's total economic potential could be between $17.1 and $25.6 trillion annually.

"Generative AI has the potential to change the anatomy of work, augmenting the capabilities of individual workers by automating some of their individual activities. Current generative AI and other technologies have the potential to automate work activities that absorb 60 to 70 percent of employees' time today." Companies across many industries are already harnessing AI to boost productivity and foster innovation and insight thanks to the ability, unlocked by AI, to manage levels of complexity and volumes of information that neither humans nor existing systems could ever hope to process.

Employers stand to benefit economically, but what about the workforce? If AI is democratized, it could be equally beneficial to individuals. AI could enhance and augment our abilities by providing everyone with a dedicated AI assistant that acts as everything from life coach to therapist to physical trainer, leading more and more people to discover fulfilling pursuits and lead satisfying lives.

Just as the Agricultural, Industrial, and Information Ages disrupted the labor and economic paradigm, as we enter the Intelligence Age, the nature of work may need to be redefined once more. Will the 40-hour work week remain the norm? With AI and eventually AIS generating substantial value and productivity gains, we may witness the emergence of new societal norms, allowing individuals more time to pursue their passions and interests, or engage in meaningful contributions to society.

AI also has the potential to revolutionize various sectors through hyper-personalization. In education, AI could analyze individual learning styles, preferences, and progress to customize educational content and methodologies for each student, leading to accelerated knowledge and skills. The combination of big data and AI could enable the development of hyper-personalized medicine, where patient data is analyzed to generate precise diagnoses, predict disease risks, and create personalized treatment plans, transforming the healthcare system. In the realm of energy management, AI could optimize energy consumption in homes and buildings by analyzing usage patterns, weather conditions, and individual preferences, resulting in reduced waste, lower utility costs, and a more sustainable future.

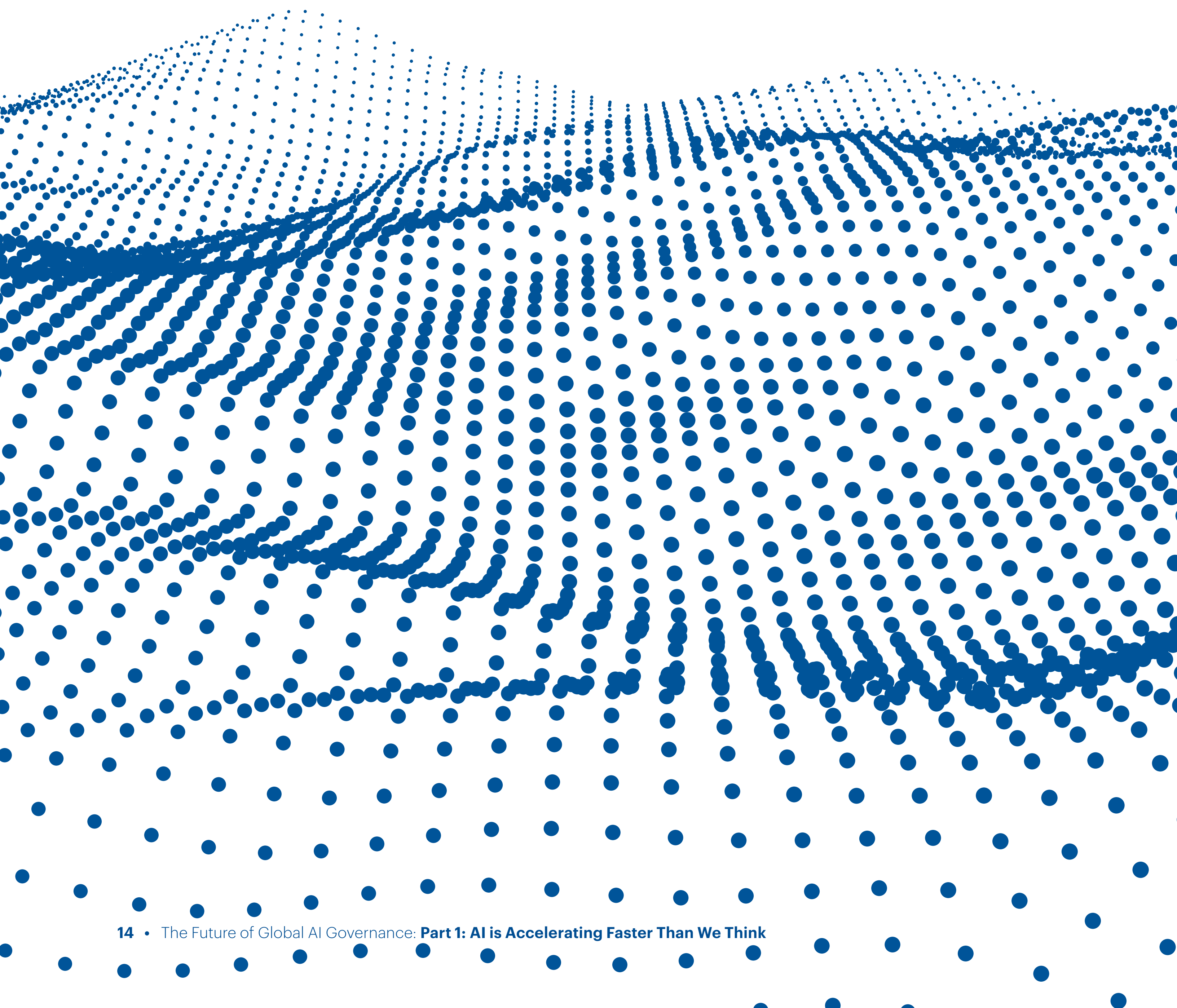
AI will accelerate scientific breakthroughs as well. For example, the AI system developed by DeepMind called AlphaFold made headlines in 2020 for its ability to accurately predict protein structures, a crucial step in drug discovery. Additionally, the World Health Organization has estimated that AI could reduce cancer-related deaths by 50% in the next 25 years.

AI-powered security systems could make us safer by analyzing data from various sources, such as surveillance cameras, sensors, and other security devices. By leveraging pattern recognition and anomaly detection, these systems could identify suspicious activities or behaviors and trigger appropriate responses, notifying security personnel, alerting authorities, or implementing automated measures to neutralize potential threats.

In the digital security realm, AI algorithms could analyze vast amounts of data that would overwhelm human cognitive abilities to detect cyber threats, such as malware, phishing attempts, and unauthorized access attempts. AI-powered cybersecurity systems could instantly identify patterns and anomalies that may indicate malicious activities, helping organizations and individuals protect their digital assets and sensitive information better than systems void of AI augmentation.

On a grand scale, AI could help us make better decisions by providing more accurate, relevant, and always up-to-date information. Greater collaboration and knowledge-sharing across cultures and geographies could encourage a more interconnected and understanding world. AI is also emerging as a promising tool to address global-scale threats like the climate crisis. Researchers are already harnessing AI to generate crisis-responsive solutions. For instance, Microsoft's AI for Earth grant focuses on building a planetary computer—an advanced geospatial search engine that aggregates data from various sources to facilitate climate decision making and prevent environmental disasters.

To steer AI toward its benefits while minimizing risks, comprehensive global AI governance is necessary. Current AI regulations are largely designed to manage AI companies or individual AI tools. However, existing AI governance methods may encounter limits if they do not account for the governance of AI as it evolves toward increased autonomy, self-improvement, and the ability to drive its own development. An innovative approach to AI alignment and regulation is required—one that directly addresses the complexities of increasingly autonomous AI systems.



PART 2

AI Governance Challenges

2.1 The State of AI Governance Today

Key Takeaways:

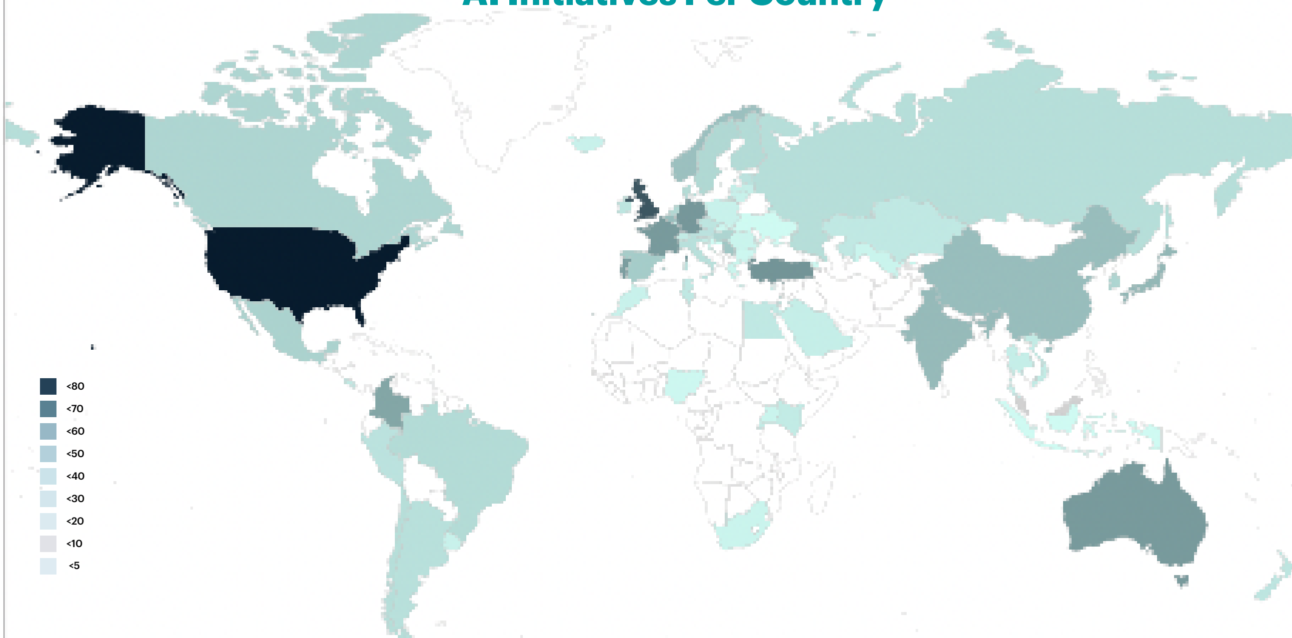
- Current regulatory and legal approaches to AI are fragmented, difficult to enforce, and largely focused on regulating people, organizations, or individual AI systems. This approach misses the bigger picture around AI networks.
- Additionally, current regulatory approaches lack coordination and agreement on methods to govern AI systems that could become increasingly autonomous and need to interact with one another.

> "Is it something that we want in our society unchecked, unverified, uncontrolled, without some clarity as to the rules and responsibilities for that sort of content? My answer and our answer in this house is no."
>
> — Dragoș Tudorache, Member of the European Parliament and co-rapporteur of the EU AI Act, speaking about the harmful content generative AI could produce.

Comprehensive and coordinated AI regulation is a necessity. How that regulation will develop, however, remains a question. There are currently over 800 AI policy initiatives in 69 countries around the world. Despite some convergence from key stakeholders on high-level principles that should govern the development and adoption of AI, including the need for transparency, accountability, fairness, privacy, and human-centeredness, there remains a lack of consensus on the most optimal path forward.

To gain a better understanding of today's regulatory landscape, we'll examine the most pertinent AI governance policies and how they have evolved, both in the United States and internationally.

AI Initiatives Per Country



OECD.AI (2021), powered by EC/OECD (2021), database of national AI policies, accessed on 7/06/2023, https://oecd.ai.

2.1.1 Initiatives in the United States

A patchwork of federal and state efforts are leading to the development of **societal standards** (e.g., fairness, transparency, non-discrimination, etc.) governing the development and deployment of AI systems.

At the Federal Level

The earliest signs of a federal AI strategy in the US were outlined during the Obama administration, including a public report issued by the National Science and Technology Council in October 2016, Preparing for the Future of Artificial Intelligence. Federal strategies and proposed AI governance frameworks have since evolved.

- The Obama administration later released the National Artificial Intelligence Research and Development Strategic Plan, updated in 2019 and 2023.
- The Trump administration signed Executive Order 13859, titled "Maintaining American Leadership in Artificial Intelligence."
- In October 2022, the Biden administration's Office of Science and Technology Policy released a Blueprint for an AI Bill of Rights with five guiding principles including (1) creating safe and effective AI systems; (2) protecting against algorithmic discrimination; (3) enhancing data privacy; (4) ensuring adequate notice and transparency; and (5) examining human alternatives.
- In June 2023, President Biden stated, "My administration is committed to safeguarding America's rights and safety, from protecting privacy to addressing bias and disinformation and making sure AI systems are safe before they are released[.]" The White House Office of Science and Technology Policy recently issued a Request for Information, seeking public input on "mitigating AI risks, protecting individuals' rights and safety, and harnessing AI to improve lives[.]"

Federal Agencies

Federal executive agencies are also taking steps to regulate the development and deployment of AI.

- In April 2022, the Department of Energy's AI Intelligence and Technology Office developed an AI Risk Management Playbook in consultation with NIST and the AI Advancement Council.
- The Department of Commerce's US Patent and Trademark Office created an AI/emerging technologies to study the impact of AI on patent and trademark examination.
- In January 2023, NIST released an AI risk management framework to help organizations better handle AI-related threats.
- In April 2023, the National Telecommunications and Information Administration launched a Request for Comment to shape a comprehensive federal government approach to AI-related risks and opportunities.
- The FTC has provided guidance on AI tools since 2020 and is focused on scrutinizing the use of generative AI that may unfairly steer individuals into harmful decisions.
- The FTC also recently issued a joint statement with the EEOC, DOJ, and CFPB addressing the use of AI and existing legal authorities, including the Americans with Disabilities Act and Title VII of the Civil Rights Act of 1964.

The Congressional and State Level

At the Congressional level, there have been a number of efforts in recent years to address AI in various industries.

- HR 3044 would amend the Federal Election Campaign Act of 1971 to provide more transparency and accountability around the use of generative AI in political advertisements.
- HR 0066 encourages Congress to focus on regulating AI in a safe and ethical manner.

- The Stop Spying Bosses Act would prohibit employers from engaging in workplace surveillance using automated decision systems.
- The American Data Privacy and Protection Act would require impact assessments around the use of AI systems if they are used in a manner that poses a "consequential risk of harm to an individual or group of individuals."
- The Filter Bubble Transparency Act would apply new requirements on platforms that use "algorithmic ranking systems," including computational processes derived from AI.
- The Consumer Online Privacy Rights Act would also regulate "algorithmic decision-making."
- In June 2023, Senator Charles Schumer (D-NY) announced the creation of the SAFE Innovation Framework to help guide Congress in developing AI regulations. As part of this process, the Senate will invite AI experts to convene a series of "AI Insights Forums" for a "new and unique approach to developing AI legislation."

Lawmakers in a number of US states have also introduced bills aimed at regulating various aspects of AI.

- In New York City, employers must now conduct "bias risk audits" of AI-enabled systems that are used to make employment decisions.
- Colorado enacted legislation in 2021 prohibiting the use of "algorithms and predictive models" by the insurance industry that unfairly discriminate based on race, gender, sexual orientation and other factors.
- Comprehensive consumer data privacy laws at the state level have also focused on AI. The California Privacy Protection Agency, for example, is considering rules to expand rights to opt-out of automated decision making under the California Consumer Privacy Act as amended by the California Privacy Rights Act (CCPA). Similar rules around profiling using automated means exist in other state privacy laws, such as Virginia, Colorado, Connecticut, Indiana, Iowa, Montana, Tennessee, and Texas.

2.1.2 Initiatives in the EU and Internationally

Regulators in Europe and in other regions are taking a slightly different approach, attempting to merge the social standards with technical standards in a framework yet to be determined.

Europe

The EU has introduced the AI Act, a comprehensive legal framework aimed at regulating different uses of AI within the EU.

- The EU AI Act specifically represents an attempt to categorize and rank various AI applications into different "risk" categories, with higher risk applications facing heavy restrictions and lower risk ones enjoying more leeway.
- The stated goal of the AI Act is to provide a "balanced and proportionate horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market."[1]
- This is no small task, as achieving broad consensus on what actually constitutes a "high risk" AI application or service has proven exceedingly difficult. In the end, whether and to what extent the AI Act becomes binding law remains to be seen.

1 EU AI Act (Draft 2021/0106 (COD)) at § 1.1.

- Given the EU's robust regulation and enforcement of new technologies in other contexts such as data privacy, the AI Act could cause a waterfall effect throughout the world in AI regulation as well.

Other Countries

Other countries are attempting to chart their own path on AI regulation.

- In the UK, the Secretary of State for Science, Innovation and Technology released a white paper, aiming to establish the UK as an "AI superpower" by presenting a framework for identifying risk while taking a "proportionate" and "pro-innovation" approach.
- In Canada, the proposed Artificial Intelligence and Data Act forms part of the country's larger effort to update its data protection laws.
- In Singapore, the National AI Strategy includes a Model AI Governance Framework.
- The Cyberspace Administration of China released its draft Administrative Measures for Generative Artificial Intelligence Services, which seeks to regulate generative AI in a manner to ensure content created by generative AI is consistent with "social order and societal morals."
- Australia and Japan are also proceeding with regulation efforts. Finally, Japan has called for a new national strategy around AI that must "plan for content and scale[.]"

2.1.3 Balanced Regulation

How does one balance between "too much" regulation, which can stifle innovation, and "too little" regulation, which can lead to a host of undesirable and unintended consequences?

Some perspective on the latter can be obtained from the mid-1990s, when the Internet first achieved mass popularity. During that time, the US passed the Telecommunications Act of 1996 ("Telecom Act") which, among other things, contained a number of provisions specifically designed to give providers of new information-related technologies substantial flexibility in deciding how new systems would be managed.

One notable example is Section 230 of the Communications Decency Act ("CDA 230") which gave providers of "interactive computer service[s]" statutory immunity from defamation and similar claims based solely upon having provided a website, blog or other online forum where certain offending statements were uttered or published. CDA 230's "hands-off" approach was considered highly controversial when first enacted, and it continues to be so today.

Even the EU's tiered "risk"-based approach does not completely solve the problem, in that too many day-to-day uses of the technology may fall into the heavily-restricted "high risk" category. Conversely, if the definition of "high risk" only covers a few extremely serious facets (e.g., using AI to concoct a virus more contagious and deadly than COVID-19), this may open the door to unchecked AI creating and/or amplifying a number of societal harms.

2.1.4 Global Collaboration

Despite the obstacles, some convergence and coordination around AI governance are beginning to take shape. In March 2023, the White House released the "National Strategy to Advance Privacy-Preserving Data Sharing and Analytics" which provides recommendations for promoting user data privacy in all fifty US states. The strategy focuses on equity and efficiency while also aiming to address ethical and socio-technical issues related to data collection and analysis, all while protecting user information.

In June 2023, the United States and Europe began collaborating on a code of conduct for AI with the goal of bridging the gap while the EU finalizes the AI Act (which could take several years, if adopted, to take effect). European Commissioner Thierry Breton stated, "We cannot afford to sit back and wait for the new law to become applicable in 2026. This is why I have started working with AI developers on an 'AI Pact' to anticipate its implementation. The AI Pact will bring together, on a voluntary basis, the main EU and non-EU actors in this field to inform and raise awareness of the principles and democratic process underlying the EU AI law. It will be fully aligned with the final text adopted by the EU co-legislators."

The collaborative efforts between the US and the EU are promising, but the effectiveness of coordination and enforcement might be hindered without a strong regulatory and AI governance framework in place.

Regulating exponential technologies that rapidly evolve and operate with increasing autonomy necessitates an unprecedented level of coordination at local, national, and international levels. Moreover, the strategic interests, goals, and capabilities are not always aligned between the US, EU, China, and other emerging countries, which will make coordination more difficult.

2.2 Explainability: A Critical Measure for Autonomy

Key Takeaways:

- Current regulatory approaches to AI include requirements for AI to be "explainable."
- Many existing AI systems may not be able to satisfy "explainability" requirements due to their inherent complexity and opacity.
- AI systems that are designed to be "explainable" and more transparent will likely have key advantages when operating under existing and future regulatory regimes.

Each of the AI governance efforts described above require a key ingredient: *explainability*. An explainable AI system has the ability to provide transparent, human-understandable explanations for its outputs. An explanation could include the factors the system considered, transparency in data sources, and the reasoning behind its predictions, decisions, and actions. An explainable system would require that stakeholders understand what happens to data going into an AI model, and how the output is generated. Explainable AI enhances trust, accountability, and comprehension of AI systems, allowing humans to understand, verify, and potentially correct the outcomes. It helps users, regulators, and stakeholders evaluate fairness, bias, and ethical implications of AI algorithms, enabling responsible and informed decision-making as it relates to regulatory compliance.

A focus on "explainability" and "transparency" of AI systems has become a key principle of "responsible" or "trustworthy" AI deployment. The Organization for Economic Co-operation and Development (OECD), for example, calls on AI stakeholders to "commit to transparency and responsible disclosures regarding AI systems," including meaningful information to "enable those affected by an AI system to understand the outcome" and "enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision."

The NIST AI risk management framework likewise says that "trustworthy" AI should be explainable and interpretable, meaning that stakeholders should be able to gain "deeper insights into the functionality and trustworthiness of the system, including its outputs."

In one proposed version of the EU AI Act, a right to "explainability" is outlined in Article 68(c) that would grant individuals the right to request a clear and meaningful explanation, in certain circumstances, on the role that the AI system played in making certain decisions, including providing the parameters and input data used when arriving at a particular decision. Although the exact parameters of the proposed Article 68(c) are still being debated, if adopted, such a right would be one of the first regulatory efforts to impose "explainability" as a technical requirement for AI systems to operate. Granting such a right would help individuals understand how AI systems are making important decisions about them, while providing individuals the ability to challenge those decisions. But this level of "explainability" may exceed what existing AI systems are able to achieve.

Within the US, states are also debating the proper scope of granting rights to stop automated decision-making, which may require a level of explainability around AI depending on how the regulations are written.

Explainability Isn't The Only Challenge

The regulatory and legal demands for explainability and transparency pose significant challenges for the generative AI industry. The decision-making process of these systems remains challenging to untangle due to their transformer-based approach. Often referred to as "black boxes," transformer-based models are complex, multilayered structures, with millions, billions, or even trillions of parameters. Their architecture makes it difficult to explain their decision-making process at all, let alone in a clear and meaningful way. Beyond their opaque design, other challenges persist:

1. **Bias and Misinformation:** Trained on vast amounts of data, including biased or inaccurate information, misinformation, or even their own generated content, generative AI systems may, if unmitigated, propagate such biases or inaccuracies in their output.

2. **Lack of Causal Reasoning:** Generative AI systems excel in identifying data patterns; however, they lack the capacity to comprehend causality. Consequently, when prompted for an explanation, these systems may generate a seemingly plausible response that aligns with observed patterns but is devoid of an understanding of the underlying causal relationships. This lack of causal reasoning can lead to flawed or misleading explanations, especially in complex, dynamic domains, such as healthcare or finance, where understanding cause and effect is critical.

3. **The Challenge of Contextual Understanding:** Although transformer-based AI models excel at producing content that can mimic cognitive reasoning to generate contextually appropriate responses, these systems lack the capacity to comprehend and reason in a natural cognitive manner. Unlike humans who possess internal representations of the world (a "world model"), transformer models rely on pattern matching. Consequently, ensuring user safety and well-being in critical scenarios, such as offering medical advice, financial recommendations, or operating a passenger vehicle becomes challenging due to limited contextual reasoning.

4. **Inadequate Handling of Novel Situations:** Lastly, these models are trained on historical data, which poses challenges when faced with novel situations not encompassed in their training data. This could lead to inaccurate or potentially harmful real-time decision-making in situations that deviate from the model's training.

The design of these models is their Achilles heel. Yann Lecun notes these systems are "insufficient to emulate the kind of learning we observe in animals and humans ... the main missing part is learning how the world works, mostly by observation without action."

Complying with regulatory requirements for explainability and transparency may require different approaches in order to achieve sufficient interpretability, fairness, and reliability. Fortunately, there are some promising research approaches aimed at improving the explainability and safety of AI systems, including:

- LIME (Local Interpretable Model-agnostic Explanations), a technique that generates explanations for individual predictions of AI systems by creating a simpler model which approximates behavior and provides insights into why predictions are made.

- SHAP (SHapley Additive exPlanations) calculates contributions of each feature of an AI system's prediction. It considers feature combinations to determine impact. SHAP explains predictions and identifies the most influential features.

- Counterfactual explanations demonstrate how predictions change through different features. By altering feature values and observing the effects, counterfactual explanations provide insights into specific predictions and highlight important features.

Each of these approaches comes with limitations. LIME can be computationally expensive and can generate difficult-to-interpret explanations. SHAP can be challenging to calculate for complex AI systems and sensitive to feature choices. Generating counterfactual explanations for complex AI systems can be difficult and interpreting them can be even more challenging.

These approaches may ultimately not meet the requirements of the EU AI Act and similar regulatory frameworks. More rigorous methods than what researchers or leading generative AI organizations have proposed will likely be needed. And though leaders in the AI industry appear eager to find a safe path forward, there are hurdles to overcome.

CEO of OpenAI Sam Altman stated his company will "try to comply" with the EU AI Act, if adopted. "Either we'll be able to solve those requirements or not," Altman said of the EU AI Act's provisions for high-risk systems. "**But there are technical limits to what's possible**."

OpenAI has since issued a white paper suggesting their technology (and similar generative models) should not be categorized as "high risk" under the EU Act's proposed framework. In a June 2023 post from a Stanford University Initiative at the Center for Research on Foundation Models, however, researchers directly assessed the ability of foundational model providers to comply with the draft EU AI Act and found that most models would struggle to align with its transparency requirements. According to the findings, foundational model providers "rarely disclose adequate information regarding the data, compute, and deployment of their models as well as the key characteristics of the models themselves." It seems that technological solutions, perhaps those that incorporate causal reasoning capabilities, may be required to address the challenges of explainability and transparency.

In a 2023 TED Talk, computer scientist Yejin Choi asked the question: **"Can AI, without robust common sense, be truly safe for humanity?"** While common sense comes naturally to humans, it remains an arduous challenge for machines. There is a clear link between common sense and causal reasoning. The ability to understand and predict cause-and-effect connections based on our everyday experiences and observations is critical. It helps guide our behavior, decision-making, and problem-solving in everyday situations. This ability may be an essential obstacle for AI systems to conquer before they can become explainable. With the ability to reason about the causes behind the data they process, AI systems could consistently and logically explain their actions and decisions. This crucial ability could be integrated in their design from the ground up, and will be explored in more detail later in this paper.

2.3 The Limitations of Current Proposed Regulations

Key Takeaways:

- Current regulatory approaches to AI should consider the long-term implications of self-regulating AI systems and ensure adaptability to future challenges.
- Regulation should focus on governing AI systems in addition to the developers and deployers of AI.
- Encoding principles, values, and guidelines directly into AI systems is crucial for creating machines that adhere to ethical and legal standards.
- Promoting interoperability and coordination among AI systems is essential for maximizing their potential and fostering effective collaboration.

Eventually all existing regulatory efforts concerning AI could encounter limits. Frameworks that rely solely on penalizing developers and operators of AI may soon become inadequate as AI systems evolve toward AIS and true autonomy, with capabilities for self-improvement and adaptation.

Consider autonomous vehicles (AVs) as an example. Currently, original equipment manufacturers convert the rules of the road into data used to set behavioral boundaries for the AI system operating AVs. However, as the number of decisions made by a growing multitude of AIs increases—and there will be potentially billions of decisions occurring every day—attempts to manage them at a human level will face limitations.

The solution could be to let AIS progressively manage more and more of these decisions. Therefore, it is crucial to seek new regulatory strategies that safely enable the increasing autonomy of AI and AIS.

Preparing For Self-Regulating and Self-Adapting Systems Today

Current regulatory approaches to AI primarily focus on present-day concerns and applications, failing to consider future implications of interconnected, adaptive, self-regulating AI systems.

In addition, certain approaches, guidelines, and frameworks for AI governance only focus on regulating the developers, corporations, and individuals behind these technologies, rather than the AI systems themselves. Although well-intentioned, this strategy is inherently limited. As AI systems progress toward higher levels of autonomy, it becomes imperative to establish frameworks that govern an AI system's behaviors and actions.

Regulatory frameworks must also factor in AI's proposed destination. Failing to account for a future where AIs network, evolve, and learn on their own could leave us unable to control outputs or outcomes, and unprepared for the challenges and complexities that advanced AI systems may inevitably bring. As these systems operate increasingly beyond our direct control, they must be equipped with the ability to self-apply various governance frameworks that preemptively encode enforcement thresholds and fail-safes. This is crucial to ensure their safety, trustworthiness, and optimal performance.

Enabling AI to Encode Our Values

> We cannot govern AI systems and the machines they will power in the same way we govern humans.

Machines do not respond to punishment, nor are they currently bound by empathy or ethical concerns. To create AI systems that will remain aligned with our values and interests, we will need a means of encoding our principles and guidelines directly into the AI software and systems. However, current proposed AI regulations lack the necessary technological definitions or mechanisms to achieve this goal, hindering our ability to establish AI systems that adhere to our moral and legal standards.

Dr. Fei-Fei Li, an AI expert and Stanford computer science professor, emphasizes the importance of considering human issues from the start when designing technical systems. This process involves integrating human perspectives into various stages, such as data acquisition, data annotation, data utilization, result interpretation, and incorporating human factors to enable conflict resolution. Dr. Li further explains "[i]f we have that collective scientific curiosity on pushing to create machines that mimic that kind of mental image of human intelligence, while at the same time understanding humans, we can make humanity better in so many ways."

Global Interoperability

Current regulatory approaches often neglect the importance of interoperability, prioritizing individual AI systems without considering the immense benefits that collaboration and coordination could bring. To unlock the full power of AI, it is imperative to establish frameworks that foster the development of AI systems capable of harmonious cooperation. Such frameworks would enable diverse AI systems to work together, transcending geographical boundaries and organizational silos.

By embracing global interoperability, AI systems could transcend their individual limitations and leverage the collective knowledge and capabilities of a networked community of AI agents. The outcome is an intelligent web of humans, AI agents, robots, and sensors. This collaborative environment enables us to address intricate challenges that exceed the capabilities of any single system. Global interoperability allows AI systems to exchange real-time data, enabling "ground truth,"and to learn from each other's experiences, making informed decisions collectively. The result is a powerful synergy that elevates the performance and impact of AI technologies. In short, global interoperability enables the web to become AI-powered.

PART 3

The Path to Autonomy and Beyond

3.1 From Autonomous Vehicles To Autonomous Everything

Key Takeaways:

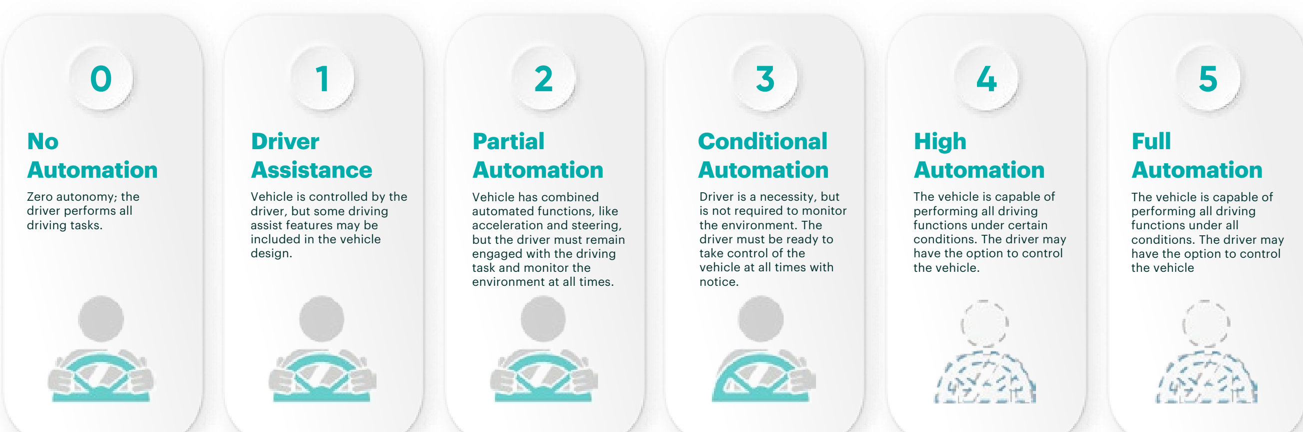
- Autonomous vehicles on our public roadways points towards the opportunities and challenges associated with AI governance of AIS becoming ubiquitous.
- The Society of Automotive Engineers (SAE) introduced a six-level classification system that provides a standard framework for categorizing the level of autonomy in self-driving vehicles, setting expectations for their performance.
- SAE levels have the potential to serve as a guidepost for a new rating system that measures autonomy levels in AIS.

While it can be challenging to comprehend the full slate of potential benefits AIS may offer, we can look to the autonomous vehicles (AVs) sector for insights into how AI may be regulated in the future.

The deployment of AVs, including cars, buses, trucks, trains, ships, and unmanned aerial vehicles (UAVs), hold great promise. Properly deployed, AVs can lead to increased efficiency and productivity, safety, reduced environmental impact, and other benefits that could reshape travel, supply chains, and even boost GDP.

In a bid to standardize this rapidly growing field, the Society of Automotive Engineers (SAE) introduced a six-level classification system for driving automation. The SAE Levels of Driving Automation categorize AVs according to their level of automation, ranging from L0, which indicates no automation, to L5, representing full automation. By providing a framework for the capabilities of AVs, SAE levels offer insight into how these machines should function in various scenarios while also establishing expectations for their performance.

Society of Automotive Engineers (SAE) Automation Levels



Source: https://www.sae.org/standards/content/j3016_202104/

SAE levels are a crucial standard for the automotive industry. They also have the potential to serve as a guidepost for a new rating system that measures autonomy levels in other sectors such as manufacturing, healthcare, and agriculture. In the upcoming sections, we explore a comparable framework for comprehending autonomy levels in industries reliant on AI systems. First, however, it is important to understand where we are in our journey toward autonomous machines, and where we are heading.

3.2 From an Analog Past to an Autonomic Future

Key Takeaways:

- The journey toward autonomous machines began with everyday automated devices and has now reached the era of autonomy in various sectors.
- Autonomic systems represent the highest stage of intelligence and autonomy, capable of independent decision-making and operation.
- The emergence of autonomic systems raises challenges in AI governance, trust, and understanding, requiring new forms of AI and governance to enable their development and integration.
- A multilevel ratings system for AIS could provide a clearer evaluation of capabilities, enhance safety, facilitate regulation, and promote innovation.

While the Intelligence Age is still in its early stages, our journey toward autonomous machines started long ago with everyday devices like automatic doors, dryers, can openers, and car transmissions. Over time, we witnessed the rise of automation, from Automated Teller Machines (ATMs) to Roombas and Nest thermostats. Now, we have arrived in the era of autonomy.

It is crucial to distinguish between "automation" and "autonomy." While automation streamlines processes, autonomy represents systems that possess their own decision-making capabilities, i.e., they have a "mind of their own."

Entrusting software with decisions traditionally handled by humans is a formidable challenge, especially when dealing with complex civic and corporate infrastructure boundaries and multiple stakeholders with potentially conflicting interests.

Nevertheless, we willingly traversed from Automatic Lane to Automated Boulevard and accelerated onto the Autonomous Highway, where we now reside in an age of autonomous cars, UAVs, delivery robots, and virtual chatbots—machines capable of independent decision-making and real-world operation.

As technological systems advance, they progress through different stages of intelligence, adaptability, and autonomy. The chart below outlines the five stages that encapsulate this evolution, showcasing the journey from analog mechanisms to future networks of autonomic AIs or "agents" that continually improve and regulate themselves. These autonomic "ecosystems" would, if realized, self-manage and self-govern, making decisions autonomously based on changing conditions and goals.

Autonomy And Beyond: Our Autonomic Future

Technology	Decision making	Complexity	Examples	System Type
Analog	None	Low	Analog clocks, mechanical devices	Single
Automatic	Rule-based	Low	Automatic doors, automatic car transmission	Single
Automated	Programmed	Medium	Assembly lines, automated email systems	Single
Autonomous	Independent	High	Self-driving cars, autonomous drones	Single
Autonomic	Self-regulating	High	Smart grid systems, adaptive traffic management	Ecosystem

Source: Spatial Web Foundation, VERSES

Gartner's Top Trends of 2022 predicts the widespread adoption of autonomic systems within the next decade, though this timeline may be accelerated given the explosion of AI in 2023. "**Autonomic systems are examples of accelerated AI automation. They are self-managing physical or software systems, performing domain-bounded tasks that exhibit three fundamental characteristics: autonomy, learning and agency.** When traditional AI techniques aren't able to achieve business adaptability, flexibility and agility, autonomic systems can be successful in helping with implementation."

While the prospect of relinquishing human control may seem daunting, drawing parallels with the autonomic nervous system's intelligent regulation of bodily functions can provide some insight. Just as our bodies strive for homeostasis (i.e., the physiological process by which organisms maintain a stable internal environment despite external fluctuations), AI-driven autonomic ecosystems would strive to regulate themselves in a similar manner.

These autonomic systems could adapt to changing conditions and optimize resources, aiming to achieve balance on local, regional, national, or even planetary scales. The emergence of such systems would rely on networks of autonomous intelligent agents working in harmony, eliminating the need for a central conductor. It is essential to emphasize this aspect as the networking of AIs remains a novel concept and is not widely discussed in the discourse surrounding the governance of these systems.

However, networks are integral to the evolution of intelligent systems. From bacterial systems to human social structures, networks operate on clearly defined standards for communication, coordination, and collaboration. Those collective systems only work if they are adaptable and coordinated. In that sense, they operate less like a classical orchestra with sheet music and a conductor to guide them and more like a jazz ensemble who know the key and time signature but are free to play and interplay off each other dynamically in real-time.

But how can we trust such systems to act appropriately on our behalf? Trust is a spectrum, and the level of trust we place in AI systems will directly affect the level of autonomy we are comfortable granting them. Trusting the behavior of an automatic door, for example, is straightforward; its actions are prescribed and its purpose is predefined—it either opens or closes. However, trusting AIS is more challenging given the complexity of the tasks these systems could undertake, such as managing hospitals, supply chains or disaster response. Ultimately, the question we must answer is this: **What is the best way to govern autonomous machines that operate most effectively when governing themselves?**

The first step is to adopt a methodology that enables humans to fully understand and therefore trust the decisions AI and AIS make. If we address the "explainability problem" and ensure that AI and AIS are transparent and accountable, we will open the door for autonomous machines to become autonomic ecosystems. Achieving this, however, may require new types of AI along with AI governance frameworks that are designed to steer these future systems before they emerge.

3.3 Steering toward Autonomous Intelligent Systems

Key Takeaways:

- AI is often presented as moving through three stages, from narrow to general to super intelligence.
- However, a more detailed perspective reveals five stages that culminate in an "ecosystem of intelligence" where multiple agents, including humans and AI systems, work together to solve complex problems and achieve goals.

> "Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, we will have multiplied the intelligence, the human biological machine intelligence of our civilization a billion-fold." - Ray Kurzweil

We don't exactly know how AI will evolve toward artificial super intelligence, but it is usually described as progressing through three stages:

1. **Stage 1: Artificial Narrow Intelligence (ANI).** These systems represent state-of-the-art AI and are designed to perform specific tasks or solve specific problems within a limited domain but are not capable of exhibiting the general intelligence found in humans. This distinction is crucial as it underscores the limitations of ANI and the need for human oversight and intervention. Types of Narrow AI include speech and image recognition software, natural language processing software, most current generative AI, and recommendation systems.
2. **Stage 2: Artificial General Intelligence (AGI).** These systems exhibit flexible, general-purpose intelligence found in humans able to adapt, learn and understand new concepts, and perform a wide range of tasks and activities.
3. **Stage 3: Artificial Super Intelligence (ASI).** These systems are an advanced form of AI that surpasses human capabilities, including those of experts, by operating at a much higher level of intelligence across various domains.

Artificial Narrow Intelligence (ANI)

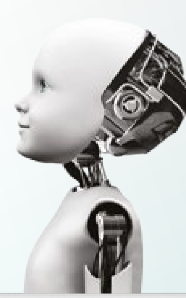
ANI describes AIs that are good at a particular task at a level equal or better than a human being.

EXAMPLE
Virtual assistants, such as Siri or Alexa.

Artificial General Intelligence (AGI)

AGI is an AI that can perform any task that a human being can. This is what most of us think of when we think of AI.

EXAMPLE
David, the child-like android from the 2001 movie Artificial Intelligence.



Artificial Super Intelligence (ASI)

This is an intelligence that surpasses anything that humans can do.

EXAMPLE
Marvel's J.A.R.V.I.S. (Just A Rather Very Intelligent System)

Source TechTalks; Vernor Vinge

In science fiction, ASI is often portrayed as a single entity, an all-knowing artificial brain. It is possible, however, that the zenith of the AI age will be a distributed network, much like the World Wide Web today, powered by "ecosystems of intelligence." In this case, the evolution of AI from narrow to super intelligence could advance through five distinct stages of the "Intelligence Spectrum."

- **I1: Systematic Intelligence.** I1 possesses the capability to identify patterns and react to them. It corresponds to the present-day advanced AI, known as Artificial Narrow Intelligence (ANI). ANI refers to machine-based software that maps inputs to outputs, optimizing a value function or cost of states. Deep learning and reinforcement learning are examples of ANI.
- **I2: Sentient Intelligence.** I2 has the ability to perceive, anticipate, and respond to the environment in real-time. This intelligence is curious and seeks both information and preferences. Such an AI would respond to sensory impressions and be able to plan based on the consequences of an action or belief about the world, which enables it to solve almost any problem.
- **I3: Sophisticated Intelligence.** I3 has the ability to learn and adapt to new situations, and thereby plan into the future. This intelligence makes plans based on the (counterfactual or imagined) consequences of an action or beliefs about the world. It moves from the question of "What will happen if I do this?" to "What will I believe or know if I do this?" This intelligence uses generative models and corresponds to AGI in the traditional sense.
- **I4: Sympathetic Intelligence.** I4 has the ability to understand and respond to the mental states, emotions and needs of people and other AIs. This type of intelligence understands the thoughts and feelings of humans and other AIs. It takes on the perspective of users and sees things from their point of view. It is capable of recognizing and understanding different perspectives concerning the same data.
- **I5: Shared (Super) Intelligence.** I5 envisions a collaborative environment where humans and AI agents join forces to tackle complex problems and accomplish objectives. This stage represents a more utopian version of ASI. It is the collective intelligence that emerges when sympathetic intelligence works together with people and other AIs; an intelligence that derives from many agents working together, creating a "living web" of shared knowledge or "an ecosystem of intelligence."

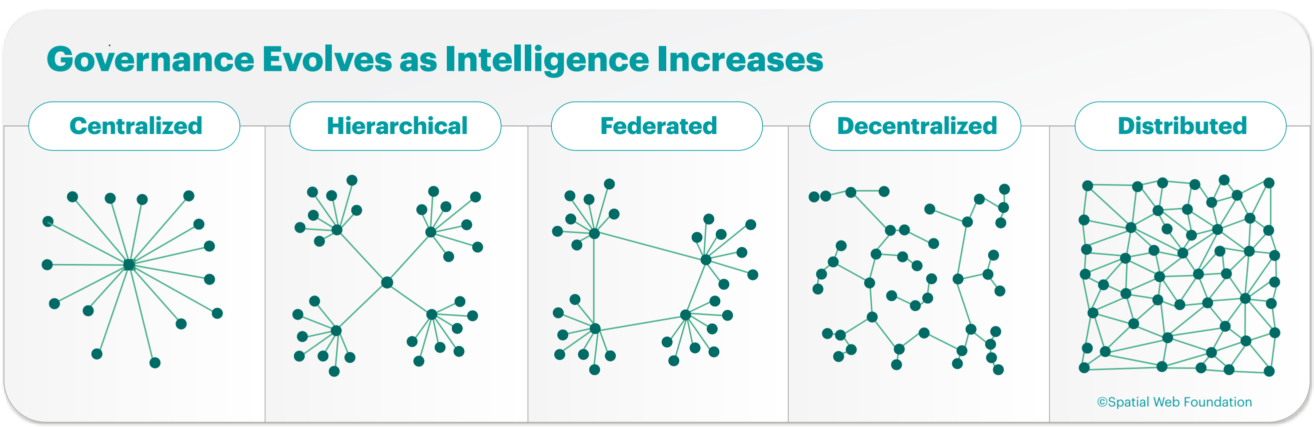
3.4 AI Governance Evolves as Intelligence Increases

Key Takeaways:

- AI governance frameworks can be applied to AIS systems based on factors such as intelligence capabilities, autonomy level, and trust.
- Potential AI governance frameworks include Centralized Governance, Hierarchical Governance, Federated Autonomy, Decentralized Autonomy, and Distributed Autonomic.

As AI systems become more intelligent, governance frameworks reflecting varying degrees of control, coordination, and distribution of decision-making authority will be required. The selection of a governance model relies on factors such as the specific application domain, the complexity of reasoning and intelligence capabilities of AIs, their potential level of autonomy, and the degree of trust we can place in them.

In other words, an AI system must attain specific levels of intelligence and adaptability within complex socio-cultural settings to be granted autonomy within particular governance frameworks. Yet the responsibility of choosing which frameworks to apply lies with regulators. As higher levels of intelligence become accessible, regulators and AI systems have various governance types to consider and potentially adopt. Here are a few examples:



- **Centralized AI Governance:** A global authority or entity exercises control and decision-making power over the AI systems. It has ultimate authority in determining the behavior, policies, and actions of the AI. This approach requires maximum trust in the centralized entity's ability to make responsible and ethical decisions.
- **Hierarchical AI Governance:** A structured hierarchy of AI systems, where high-level entities including humans and AIs have control over lower-level AIs. Decision-making flows from the top down, as higher-level entities provide guidance and instructions to the lower-level AIs. Trust is placed in the hierarchical structure and the competence of the higher-level entities.
- **Federated AI Governance:** A collaborative network where multiple autonomous AI systems either form a consortium or operate under federated principles to work toward shared objectives. Decision-making is distributed among members, balancing individual autonomy with collective concerns. Trust is

placed in the collective intelligence, cooperation, and shared responsibility of the AI systems. Each member contributes its expertise and decisions are reached through consensus, adhering to the overall governance framework and fostering a cooperative and responsible environment.

- **Decentralized AI Governance:** In decentralized systems, AI systems operate autonomously, with minimal central control or coordination. Each AI system possesses a high level of autonomy and makes decisions based on local information and knowledge. Trust is distributed among the individual AI systems, emphasizing their ability to act independently and responsibly, and in the best interest of their localized domain.
- **Distributed AI Governance:** This level of AI governance is characterized by a network of autonomous AI systems with direct connections that self-regulate and adapt to multiscale conditions and contexts in real-time. Trust in this "autonomic" network governance is implicit, residing in the network's ability to self-regulate, adapt, and make context-aware decisions from local to global scales. It incorporates elements of centralized, hierarchical, federated, and decentralized governance while maintaining the flexibility to respond to both local and global dynamics. Distributed networks ensure safety, efficiency, and adaptability, fostering collective adaptation based on shared knowledge, ultimately embracing the dynamism of a fully autonomic world.

The combination of increasingly intelligent AI and new types of governance raises a critical question:

> How can we determine the appropriate level of autonomy—i.e., what governance type is appropriate—to assign to AI systems in order to maximize potential and minimize risk?

In the following section, this question is addressed with a proposed multilevel ratings system similar to the SAE levels used for AVs that could provide a comprehensive framework for evaluating the capabilities and limitations of AIS. This approach would offer a more precise, responsible, and nuanced assessment, benefiting users, developers, and regulators alike.

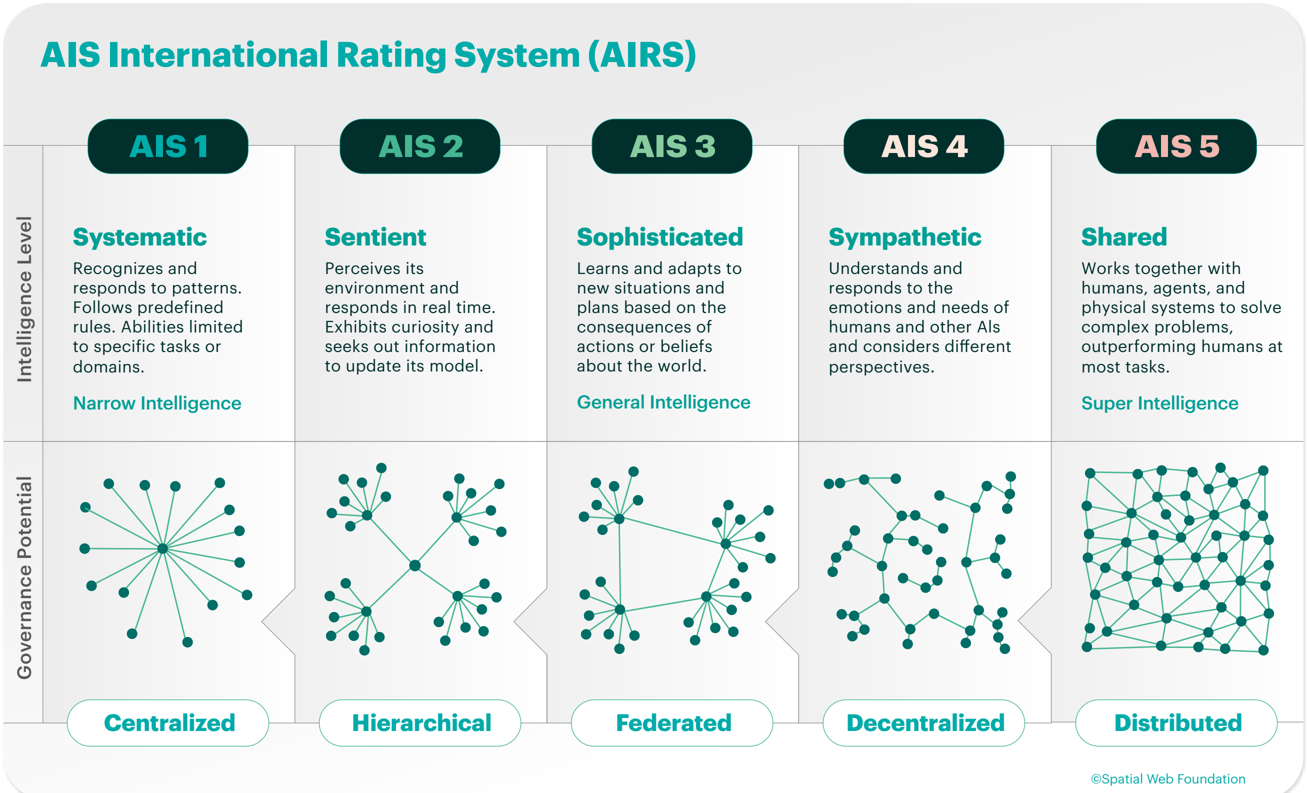
3.5 AIS International Rating System (AIRS)

Key Takeaways:

- A multilevel AIS International Rating System (AIRS) akin to the SAE levels for AVs could be applicable to all AIS.
- The level of governance that is possible, ranging from centralized to distributed, is determined by the level of intelligence of AI and potentially the level of intelligence of a network of AIs, or AIS.
- AI systems networking across diverse domains and industries necessitates the development of standards to facilitate seamless interaction, understanding, and governance between networks.

By merging the underlying concept of a multilevel classification system for the automation of vehicles as outlined in the SAE 5, combined with the intelligence stages described in the "Designing Ecosystems of Intelligence from First Principles" research paper, we begin to see the outline for a mulitlevel AIS International Rating System (AIRS).

The chart below provides the thresholds for AGI and ASI, respectively, and depicts the correlation between autonomy, intelligence levels, and the types of possible AI governance structures unlocked at each stage.



As AI systems become more intelligent, they gain the potential for greater autonomy, which would be reflected in the corresponding governance framework that becomes available, along with all other governance frameworks that came before it.

For example, an intelligent system at AIS 4, referred to as "Sympathetic," can function within a centralized, hierarchical, federated, or decentralized architecture. However, it is expected that an AI system at AIS 2 will be limited to centralized or hierarchical frameworks due to its lower level of intelligence. It is important to note that no specific governance framework is inherently superior to others. They can be viewed as a measure of the autonomy an AI system may be granted given its level of intelligence.

Ultimately, intelligent machines may operate optimally across a range of governance frameworks. To enable this spectrum, establishing standards that facilitate interoperability among these systems will be essential.

Here is a snapshot of what could be expected at each stage with each type of AIS, using AVs to illustrate what is possible:

AIS 1: Systematic Automated Systems

(Threshold for ANI)

- These systems operate based on rule-based algorithms or machine learning models. They are limited to specific tasks and lack the ability to perceive their environment or make complex decisions.
- **Example:** Basic self-driving cars that can maintain a steady speed and stay within a designated lane on the highway. AVs follow predefined rules to automate specific tasks, such as lane-keeping or adaptive cruise control.
- **Possible governance structure:** Centralized Governance. Given their rule-based operation, these systems require centralized governance, where a central authority ensures responsible and ethical decisions. Current regulatory discussions are generally focused on this type of governance—regulating the organizations and people developing and deploying AI.

AIS 2: Sentient Automated Systems

- These systems can perceive their environment and respond in real-time. They exhibit curiosity and can seek out information, and have real-time data processing and simple decision-making capabilities.
- **Example:** An AV that can perceive its environment in real-time, process data, and make decisions accordingly is able to change its speed and driving style when it encounters a hail storm. If it meets a weather event it doesn't understand, it can seek out information to improve its driving capabilities.
- **Possible governance structure:** Hierarchical Governance. As these systems can perceive their environment and respond in real-time, a hierarchical structure can provide guidance and oversight.

AIS 3: Sophisticated Autonomous Systems

(Threshold for AGI)

- These systems can learn and adapt to new situations independently, marking the onset of AGI. They can make plans based on the consequences of actions or beliefs about the world, employing advanced frameworks like Active Inference.
- **Example:** An AV that can learn from its experiences, adapt to different road conditions, and make informed decisions, optimizes its fuel efficiency while also minimizing the time spent in traffic due to an accident up the road.
- **Possible governance structure:** Federated Governance. Because these systems can learn and adapt independently, federated governance strikes a balance between individual autonomy and collective decision-making.

AIS 4: Sympathetic Autonomous Systems

- These systems can understand and respond to the emotions and needs of humans and other AIs. They can consider different perspectives, including those of other AIs.
- **AV use case:** An AV ride-sharing vehicle can detect the mood of passengers through facial recognition and adjust the cabin environment (e.g., lighting, temperature, music) to create a comfortable and enjoyable experience. Such vehicles possess empathetic capabilities, enabling a more personalized and human-like driving experience.
- **Possible governance structure:** Decentralized Governance. Given the ability to understand and respond to the emotions and needs of humans and other AIs, decentralized governance enables individual systems to act independently, in line with their localized understanding and empathy.

AIS 5: Shared Autonomic Systems

(Threshold for ASI)

- This level represents ASI, where the AI becomes part of a comprehensive ecosystem of interconnected AI software and Cyber-Physical Systems. These systems are capable of outperforming humans at most economically valuable work. They can work together with humans, other agents, and physical systems to solve complex problems.
- **Example:** A network of AVs work together in a smart city environment to optimize traffic flow, reduce congestion, open pathways for emergency vehicles, and minimize travel time by leveraging real-time data and predictive algorithms.
- **Possible governance structure:** Distributed Governance. At this level, AIS operates as part of an interconnected ecosystem. Distributed governance allows these systems to self-regulate and adapt to conditions and contexts in real-time, with the ability to make decisions from local to global scales.

It is worth noting that the highest level of AI governance achieved by humans today is AIS 3, federated governance. It is possible with the assistance of AI that we may collectively achieve AIS 4 or AIS 5 within a generation.

As we progress beyond AIS 3, new possibilities await that could have a transformative impact on the future. Let's consider then the potentially beneficial effect that higher levels of AIS could have on societal governance overall.

What Future Governance of AIS Might Look Like

As AI systems evolve toward AIS 4, Decentralized *Autonomous* Governance becomes available. Using the principles of decentralization, AIS can pave the way for an AI governance model that distributes power among a wide array of independent nodes or actors. In this scenario, each participant, including developers, users, and policymakers, would have the capacity and authority to make decisions based on local information and knowledge.

A decentralized system has the potential to empower local communities and individuals, making AI governance more democratic. Each person, empowered by personal AIS tools, could contribute to decision-making processes, creating a truly participatory democracy.

This could also lead to innovative solutions for problems that are best tackled at and tailored to the local situations and needs, yet coordinated on a larger scale, ensuring that the unique socio-cultural dynamics of each locality are taken into account. For instance, climate change could be addressed through localized green initiatives, yet coordinated globally to ensure a comprehensive approach.

Eventually, AI systems will evolve to AIS 5, Distributed *Autonomic* Governance, which potentially represents the most advanced level of AI governance, characterized by a network of autonomous AI systems that self-regulate and adapt to multiscale conditions and contexts in real-time.

In this future society, advanced AIS would allow seamless coordination of complex systems at both local and global scales. Each node or participant in this distributed system regardless of their socio-cultural context, would not only be capable of independent decision-making but would also be able to adapt and respond to changes in real-time, creating a truly dynamic, responsive, and culturally sensitive governance system.

Imagine a global response to a socio-economical, health, and climate crisis in such a society. Real-time data from every part of the globe would be collected and analyzed, with each local system making decisions based on the global information, while also contributing local data and insights. This would allow for a highly responsive and efficient global response, ensuring that resources are allocated where they're needed most.

Such advancements could lead to a society where AI governance is dynamically responsive, context-specific, and adaptive. A new era of decentralized and distributed decision-making could empower individuals and communities and foster global coordination and cooperation. However, it is important to acknowledge that the phases and advancements outlined above may not occur simultaneously nor uniformly across regions and industries, due to varying socio-cultural contexts, economic conditions, and technological capabilities. The speed at which AI progresses and the adoption of corresponding governance frameworks will vary. This raises an important question:

> How will we manage the compounding complexity of interconnected networks of AI systems that operate across different domains, industries, and locations, with different levels of autonomy?

3.6 Governing Networks of Complexity

Key Takeaways:

- Different types of AIS may evolve at varying speeds, with distinct trajectories and emergent capabilities, suited to the context in which the AIS operate.
- To govern increasingly autonomous AI systems, we must navigate the interplay between various types of AI systems, their autonomy levels, and the governance frameworks they operate within.
- Socio-technical standards empower us to determine the extent of human involvement in the governance of AI and AIS, enabling effective steering regardless of the technology's direction.

As expressed in the previous section, different governance types will be unlocked as AI systems become more intelligent. However, not all regions or governing bodies will necessarily make use of the same governance levels at the same time, nor will they share the same cultural norms and values, which presents yet another challenge as we move gradually toward a "network of networks."

Adaptable AI Governance

As AI technology progresses and becomes more pervasive, it is essential to establish AI governance frameworks that can be adapted to the diverse cultural, legal, and ethical contexts across different regions.

While a global "code of conduct" is important, tailoring AI governance to regional variations allows for a more nuanced and inclusive approach. However, it is important to note that the adaptability of AI governance frameworks does not dictate ethical standards but serves as a tool to implement values and principles aligned with regional perspectives. By considering specific regional regulations, ethical frameworks, and societal norms, AI governance can effectively address the unique challenges and opportunities emerging worldwide.

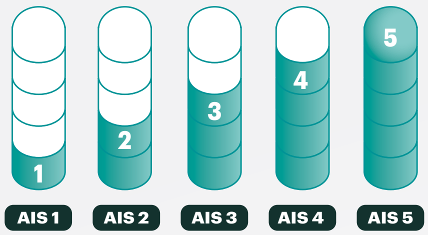
A Spectrum of Intelligence

The emergence of interconnected networks of intelligent systems is inevitable. According to Metcalfe's law, the value and usefulness of a network grows exponentially as the number of connected users or devices increases. The implications of evolving networks of intelligence, the emergence of novel types of AI, and their varied rates of development could make it difficult to steer AI in the right direction.

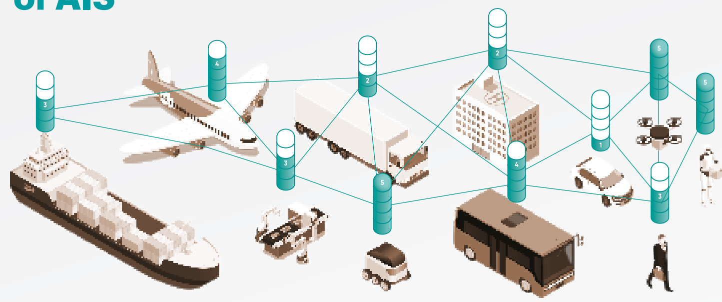
Different types of intelligent systems may evolve at varying speeds around the world. The journey toward AGI and ASI could very well take place within the next two decades, with distinct trajectories, timelines, and emergent capabilities for different AI approaches, such as Language Models and Active Inference systems.

While some AI systems may rapidly progress toward higher levels of sophistication, others may advance at a more gradual pace. It is essential that these diverse AI systems interoperate and work together in the future despite their ability to learn and adapt. As systems evolve toward these capabilities, it is imperative that AI governance frameworks can support this ever-evolving spectrum of intelligence. That means that different levels and types of AIS must be able to form networks, as in the graphic below:

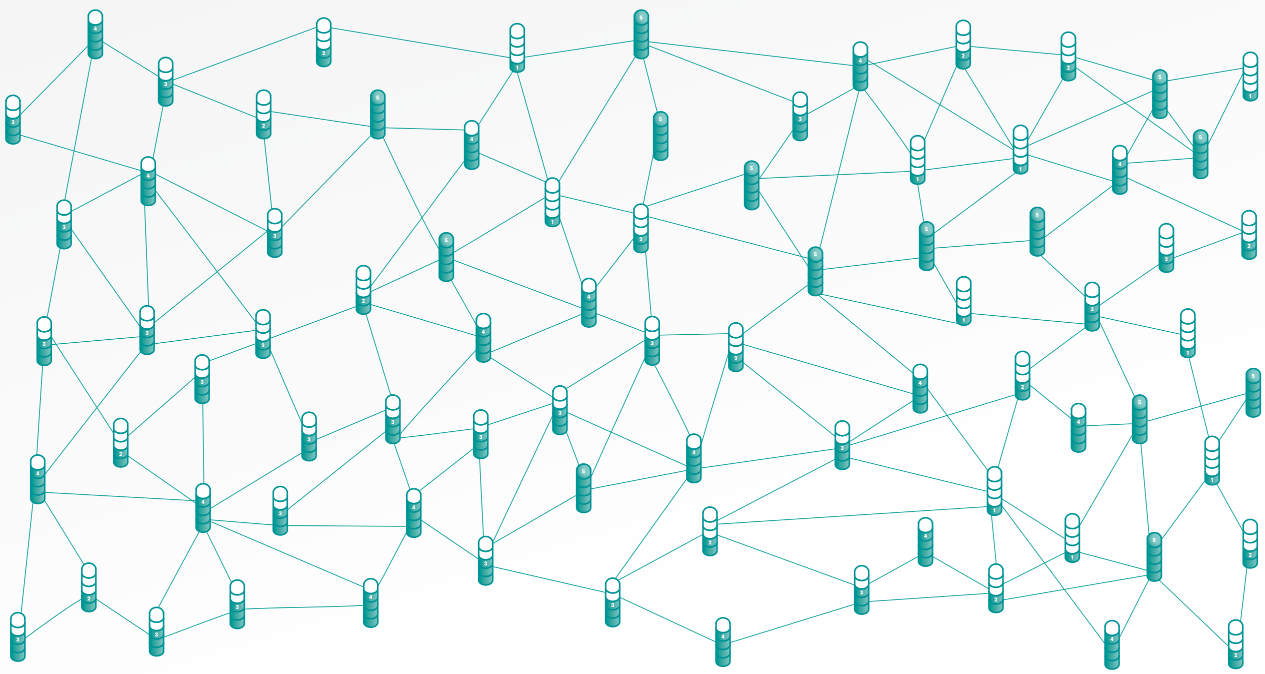
Different Levels of AIS



Different Types of AIS



Different Levels and Types of AIS in a network of Distributed Governance



Source: Spatial Web Foundation, VERSES

Dynamic Governance

AI systems are expected to operate in diverse environmental settings, adapting to and influencing their surroundings. The governance of these ever-evolving networks of complexity demands a holistic and adaptive approach.

As stated previously, the AI governance challenge extends far beyond individual AI systems. In fact it encompasses the compounding complexity of networks, where interconnected AI systems interact, learn, adapt, and include and interact with humans, further highlighting the need for advanced robust governance mechanisms.

In order for AIS to be widely adopted, a significant level of acceptance among stakeholders and regulators is necessary. At the core of acceptance lies trust. Building and maintaining trust in AIS is contingent upon the establishment of robust governance frameworks. However, it is important to note that governance frameworks may not be uniform; they will likely exhibit heterogeneity, ranging from centralized to federated to autonomic—perhaps all in one network. The choice of governance level(s) may vary depending on the jurisdiction and its specific considerations. For example, a country like China may opt for centralized and/or hierarchical governance, even with AIS operating at an average autonomy level of 4, whereas the US may choose federated or decentralized governance at the same autonomy level.

The diversity of governance frameworks adds another layer of complexity to managing AIS interoperability and highlights the importance of tailoring governance approaches to meet the specific needs and values of different regions and societies.

Trust Continuum

As AI systems become more advanced and collaborate with one another to solve complex problems, there is a growing need for a "trust continuum" to ensure transparency and accountability. This continuum is a complementary concept that applies to the governance of AI systems as they work across borders that have different regulations and laws.

For instance, different countries have different laws regarding the collection, storage, and use of personal data. The different AI systems involved will need to comply with all applicable regulations in the region where they are active.

Another concern is intellectual property rights. In a chain of multiple AI systems, there may be issues regarding the use of intellectual property and data. Maintaining a "chain of governance" presents a complex set of regulatory challenges that must be addressed to ensure global interoperability between AI systems working across borders. Without a solution, AI systems will be like drivers passing in the night, blindfolded, unable to communicate or understand the dependencies or interactions with their respective environments.

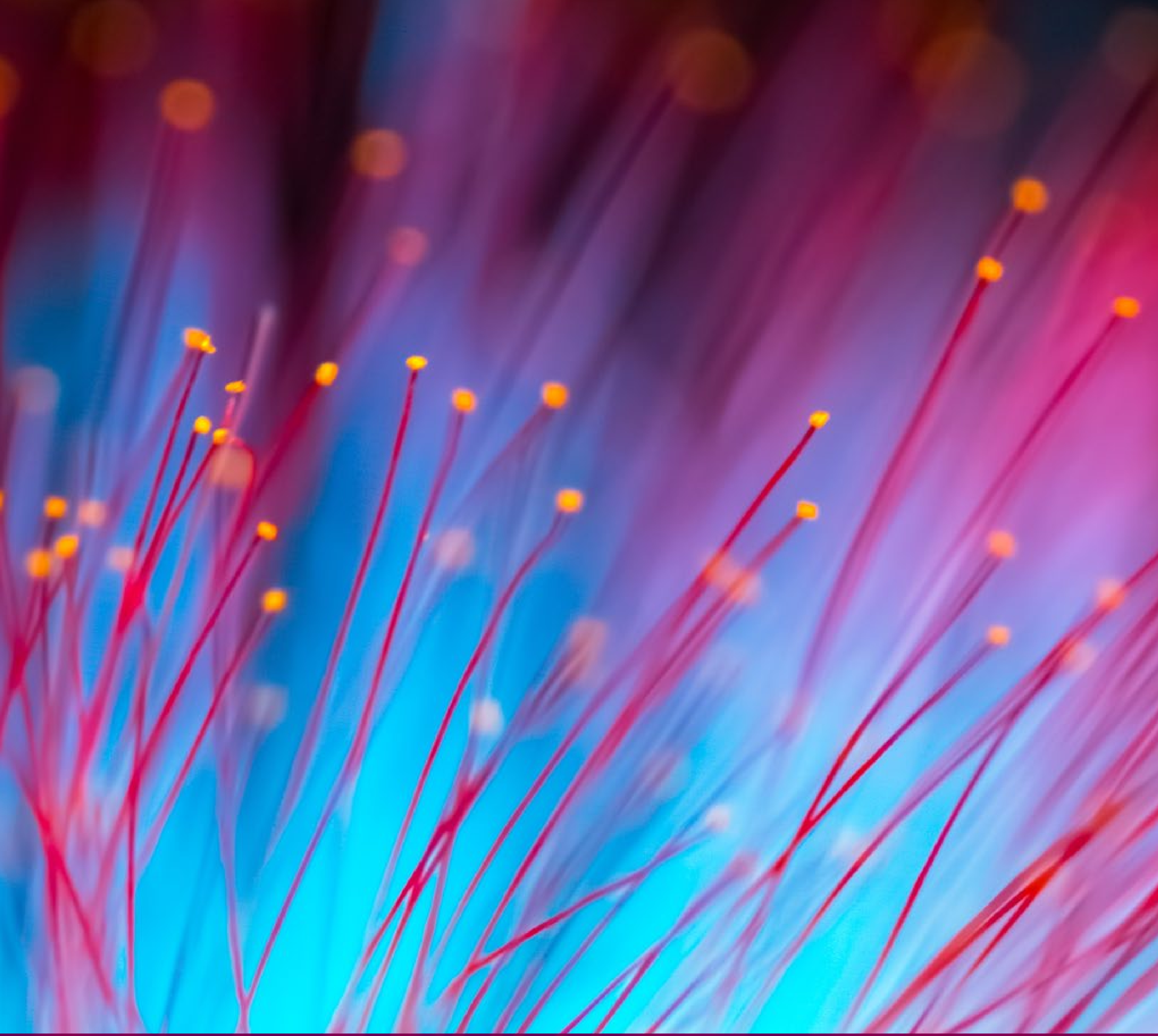
The level of trust required will vary based on the nature of the task and the potential consequences of failure. Thus, a "chain of explainability" may also be necessary. For example, in high-risk scenarios such as an AI piloting a train carrying toxic chemicals, an AI system that is one hundred percent explainable may be essential to quickly identify and rectify any errors. In less critical scenarios, such as an AI recommending a movie or product to a user, the need for a system's complete transparency may not be as urgent. Additionally, explainable AI systems may utilize other AI systems with less explainability to complete tasks. End users will need to be able to verify the level of explainability of each AI system utilized at every step in the process.

Humans in the Loop

To address the question of human involvement in the governance of AI and AIS at each AIS level, the development of **interoperable socio-technical standards** becomes paramount. Such standards, coupled with the emergence of adaptable, self-learning AI, could empower us to determine the extent of human oversight required as the technology evolves. As intelligence levels increase, such as in AIS 4 or AIS 5, the possibility arises for humans to step further out of the loop.

> By establishing a spectrum of human involvement through socio-technical standards, we can navigate the unpredictable nature of AI development and ensure effective steering regardless of the direction AI takes.

In the following sections, we will explore how socio-technical standards may be necessary to determine the level of explainability required for each task, allowing for safe and effective collaboration between humans and AIS. We will examine the structure of these standards and how the combination of an interoperable standard and adaptive AI could provide the flexibility and resilience needed to maintain human control over AI while harnessing its potential for progress.



PART 4

The Power of Standards

4.1 Standards Shape Our Modern World

Key Takeaways:

- Standards promote seamless integration, compatibility, and collaboration between different systems, products, and devices, and benefit manufacturers, regulators, consumers, and businesses.
- Standards play a vital role in various industries and have a transformative impact on safety, efficiency, and accessibility.
- Establishing a technical standard for AIS and human-machine communication is crucial to harness the benefits of AI while protecting society.

> "Our modern existence depends on things we can take for granted. Cars run on gas from any gas station, the plugs for electrical devices fit into any socket, and smartphones connect to anything equipped with Bluetooth. All of these conveniences depend on technical standards, the silent and often forgotten foundations of technological societies."
> - The New York Times

Before we explore the concept of socio-technical standards, it is important to understand the pivotal role that purely "technical" standards have played in shaping our modern world.

Technical standards provide significant benefits for governments, developers, manufacturers, and consumers alike, permeating every aspect of our modern lives. From enabling phone calls to filling up cars at the gas station to powering devices plugged into walls, standards help to ensure seamless communication, connection, and collaboration.

With tens of thousands of active technical standards in various industries and sectors, they have become the cornerstone of progress and innovation. Standards are responsible for fostering interoperability, compatibility, and consistency between different systems, products, and devices. They serve as a crucial foundation for seamless integration and collaboration, allowing different technologies to work together seamlessly.

Technical standards are documented agreements that contain precise criteria and specifications, serving as guidelines and definitions of characteristics that ensure materials, products, processes, systems and services are fit for their purpose.

They provide a common language and specifications that promote interoperability, consistency, safety, and efficiency across industries and technologies. Consider the increased efficiency that the standardization of shipping containers had on the global supply chain, allowing for seamless transfer of goods across different modes of transportation, simplifying the logistics processes. By employing technical standards, businesses and governments can work together to ensure a more consistent and reliable approach to products, services, and processes.

Technical standards are useful because they:

- **Ensure compatibility and promote global interoperability** between products and services, facilitating seamless integration and communication.
- **Encourage public-private partnerships** by enabling companies to keep certain information proprietary while also sharing the data necessary to maintain innovation.
- **Improve safety** by providing guidelines that reduce risks and hazards.
- **Promote efficiency** by streamlining processes and reducing redundancies.

- **Enhance product quality** by establishing minimum performance criteria.
- **Facilitate global trade** by harmonizing regulations and removing technical barriers.

How stakeholders could benefit from technical standards:

- **Manufacturers** can achieve economies of scale, reduce production costs, and access new markets.
- **Regulators** can use standards as a basis for legislation and ensure compliance with safety and quality requirements.
- **Consumers** can trust that standardized products and services are safe, reliable, and compatible with their existing systems.
- **Businesses** can increase their competitive advantage by adhering to industry best practices.

Organizations such as the International Organization for Standardization (ISO), American National Standards Institute (ANSI), Institute of Electrical and Electronics Engineers (IEEE), World Wide Web Consortium (W3C), and the European Committee for Standardization (CEN) have developed standards that have revolutionized communication, trade, and information exchange. These standards-setting organizations have consistently showcased the transformative impact of well-defined standards in enhancing safety, efficiency, and accessibility across interconnected public and private sectors. The table below gives a high-level overview of some of the main standardization bodies and their purpose.

Examples of Standards Organizations

Organization	Acronym	Purpose	Scope
International Organization for Standardization	ISO	Develop global standards for various industries, ensuring quality, safety, and compatibility	International
Institute of Electrical and Electronics Engineers	IEEE	Implement standards for technologies like Wi-Fi, advancing communication and information access	International
Internet Engineering Task Force	IETF	Responsible for the technical standards that make up the Internet protocol suite (TCP/IP)	International
World Wide Web Consortium	W3C	Develop web standards (HTML, CSS) to enable global communication and information exchange	International
Open Geospatial Consortium	OGC	Set standards for geospatial data integration across applications and services	International
European Committee for Standardization	CEN	Develop standards for various domains to enhance safety, efficiency, and accessibility	European Union
The European Telecommunications Standards Institute	ETSI	Develops standards for telecommunications, broadcasting and other electronic communications networks and services	European Union
American National Standards Institute	ANSI	Oversees development of standards for products, services, processes, systems, and personnel	United States
British Standards Institution	BSI	Develop standards for multiple industries in the United Kingdom	United Kingdom

Source: Spatial Web Foundation, VERSES Research

By design, standards are neutral and apply to all stakeholders. They are therefore attractive in the AI space because they may steer us in a new direction that transcends the limitations of traditional market and government approaches to AI governance.

In April 2023 at the G7 Hiroshima Summit, the Digital and Tech Ministers' meeting report stated "In 2020, we supported the launch of the Global Partnership on Artificial Intelligence (GPAI). The OECD AI Principles provide guidance for trustworthy AI and for ensuring an open and enabling environment for AI development and deployment that is grounded in human rights and democratic values... Rapid AI developments call for attention to, and cooperation on, emerging and medium-term policy issues including the development of technical standards, developed by international Standards Development Organisations (SDOs), as well as other tools to ensure the development and deployment of trustworthy AI in line with the OECD AI Principles."

In the May 2023 G7 communiqué, leaders recognized that "governance, public safety, and human rights challenges should be addressed at the global level. We task our relevant Ministers to consider collective approaches in this area, including in terms of interoperability, portability, and standards..." They further "recognize[d] the need to immediately take stock of the opportunities and challenges of generative AI, which is increasingly prominent across countries and sectors, and encourage international organizations."

The need and urgency for standards for AI and AIS that can provide clear reliable guidance for manufacturers and developers, and enable international regulatory consensus while accounting for regional variations (much like power adapters for electricity today) is clear. Forethought is essential. Attempting to add comprehensive standards to a system after the fact has historically been an unfruitful endeavor.

Take for example the state of Web 2.0, with its inherent privacy and security flaws, along with the negative side effects of attention-mining social media and misinformation stands as a prime example. Vint Cerf, the co-creator of the Internet's protocols, including TCP/IP has stated that he "didn't pay enough attention to security" and wishes he had "put more end-to-end security in the system to begin with." Today's online fraud, data theft and industrial cybersecurity risks are, in the end, the result of this lack of foresight.

The Anti-Human Web

In a 2018 interview, Sir Tim Berners-Lee, commonly viewed as the inventor of the World Wide Web, expressed disappointment with how, in the Web 2.0 era, his creation has been distorted and manipulated into something that he believes is harmful to humanity. "I was devastated," he said. "We demonstrated that the web had failed instead of served humanity, as it was supposed to have done, and failed in many places." Berners-Lee argues that centralized control unintentionally led to an "anti-human" phenomenon, resulting in a web that can be used as a tool to harm society rather than aid it.

4.2 Grounding AI: Lessons from the Electrical Age

Key Takeaways:

- The introduction of electricity led to outsized economic growth, but the potential impact of AI on collective intelligence is even greater.
- The early years of electricity were marked by uncertainty and apprehension due to accidents and lack of safety standards, similar to the concerns surrounding AI today.
- Technical standards, such as electrical grounding standards, were crucial in mitigating risks, alleviating fears, and driving global productivity. Drawing inspiration from the grounding standards of the electrical age, new standards for AI could provide essential guidelines for its safe and effective deployment.

Andrew Ng, founder of DeepLearning AI and a pioneer in the field, famously declared that "AI is the new electricity" claiming it will "transform every industry and create huge economic value."

In the early 20th century, the introduction of electricity spurred a period of outsized economic growth across all fields, from transportation to communication to manufacturing. The GDP of the US alone grew nearly tenfold compared to the pre-electrification era.

Capturing lightning in a bottle changed the world. But the power of electricity pales in comparison to the potential impact of AI. Adding electricity to a system amplifies its power, but adding AI to a system augments it with intelligence.

Like electricity, the promise of AI has also been tempered by uncertainty and apprehension reminiscent of the early years of electricity, when live current was carried through bare copper wires with minimal insulation and no grounding. Unsurprisingly, this resulted in devastating, sometimes fatal accidents. Thomas Edison's campaign against alternating current along with public electrocutions of animals heightened this anxiety.

As a result, the public, fearing electricity, shunned its development and deployment. Medical professionals, labor unions, and local communities raised concerns about health risks, job losses, and environmental impacts. But the issue was not with electricity itself, rather it was the absence of technical safety standards that created risk.

Only when measures such as grounding standards were adopted did fears around electricity alleviate and productivity and global GDP soar. Specifically, it was the IEEE that eventually created the grounding standards upon which we've built our modern technological society. They standardized the electrical grid and established uniform specifications for electrical machines, including the acceptable voltage levels they could handle. The result was the compatibility and interchangeability between various devices and systems. In short, by making electrical transmission and use significantly safer, grounding standards played a pivotal role in facilitating the electrification of the world.

As AI endures its own saga of skepticism, critique, and existential dread, we might look to the grounding standards of the electrical age for inspiration. Standards provided essential guidelines for the safe deployment of electricity and a new type of standard could do the same for AI. The first, and most crucial step in grounding AI is to enable a shared understanding between humans and intelligent machines.

4.3 Shared Understanding Between Humans and AI

Key Takeaways:

- Bridging the communication gap between humans and AI requires a shared understanding and common language.
- Gradually reducing human control over AI while maintaining oversight is crucial as AI becomes more sophisticated, allowing for course correction and shared understanding.
- International consensus on standards is necessary to ensure safe and trustworthy AI integration, with a focus on establishing a shared representation of the world and seamless interaction across network components.

In 1992, the World Wide Web arrived, revolutionizing human communication. Roughly fifteen years later, in 2007, Web 2.0 emerged with the birth of the smartphone, and an explosion of content enablement and curation followed. Fifteen years after that, in 2022, AI burst onto the scene with its newfound powers of content creation and context awareness. This, along with other technological developments such as augmented and virtual reality, mark a turning point in what is considered the dawn of the Web 3.0 era.

Leading futurists, thought leaders, and organizations such as Deloitte refer to this next-generation of the internet as the Spatial Web because of its emphasis on how computing moves out from behind our screens into the world around us. This "Internet of Everything" presumes the integration and formation of a global-scale network made up of billions of sensors and robots, user-owned data, and immersive experiences—all powered by AI.

The Spatial Web becomes the network infrastructure that AIs will use to connect to each other, to the physical world, and to us.

> The most pressing challenge will be to align AI with our agreed-upon principles. To do that requires bridging the understanding and communication gap between humans and AI.

Humans express their understanding of the world in words, while machines operate on bits. For any successful agreement to be formed, both parties must be able to understand the terms and conditions, which is impossible if they cannot converse in a way that allows for accurate interpretations of meaning. If the problem is interpretability, the solution, it seems, lies in translation. A translation tool that facilitates bidirectional conversation is needed to establish trust.

Successful translation requires an interpretation of the subtleties and nuances of context. Establishing new tools for modeling, translation, and interpretation could bridge the gap between the different human and machine worlds, enabling us to make adjustments and updates when necessary. It would also allow us to establish enough trust to let machines operate on their own, because these machines would possess common sense, which computer scientist Yejin Choi pointed to as a necessity. Think of it like a student-teacher relationship, where the teacher instills intellectual tools and life lessons, but also allows the student to learn from their mistakes. As the student matures, the teacher intervenes less and less in their daily instruction.

Similarly, as AI becomes more sophisticated, we can gradually reduce our control while maintaining the ability to exercise oversight and override measures. **AI will inevitably make mistakes, and so will we. It is crucial that we can course correct as we go. However, this requires bidirectional communication and a shared understanding of the human world.**

In order to enable the safe and trustworthy operation of countless AIs working together with humans in our streets, homes, schools, institutions, and across the globe, we must achieve international consensus on the standards that define the contexts under which AIS can operate. These standards must be sufficiently generalizable to accommodate a wide array of situations while remaining specific enough to apply to unique regional and categorical domains.

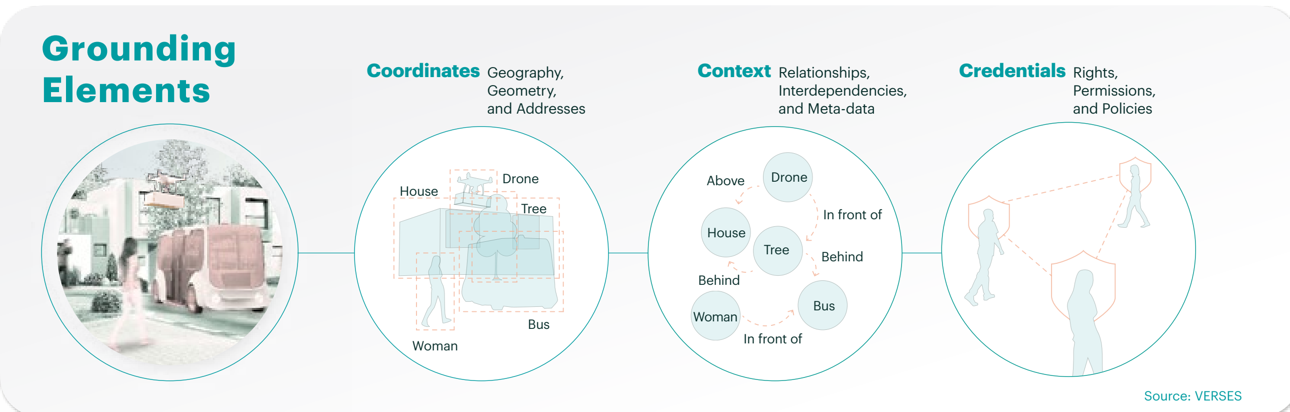
More specifically, for a network to effectively be in synchronicity, all of its nodes need to share the same representation or "model" of the world they are operating in. To accomplish this, we need a new generation of "technical" standards that facilitate seamless interactions and a shared understanding of the context of any given use case.

Defining a common "language" of operation and interaction through a shared world model would foster collaboration and integration across industries, sectors, and borders. This *lingua franca* would act as a 21st century Rosetta Stone, facilitating interoperability and understanding between different AIS, whether they're software applications, IoT devices, robots, or drones.

4.4 Grounding AI Systems

Key Takeaways:

- Grounding AI involves creating internal representations that accurately describe real-world entities, actions, and relationships.
- Knowledge Representation and Reasoning (KRR) is a foundational aspect of AI research that enables AI systems to understand and reason about their complex environments.
- Grounding facilitates meaningful understanding, collaboration, and ethical decision-making by bridging the gap between AI systems and humans, ensuring alignment with societal norms and values.



> The concept of grounding AI involves creating internal representations, known as "world models," within AI systems that accurately describe real-world entities (people, places, things), the actions that they take, and the relationships between them.

Think of grounding AI as similar to teaching a child how to interact with the world around them. Just as a child learns about people, places, and objects, and how they interact, AI systems must construct internal representations of entities in the world and their relationships to other entities or concepts.

This process enables AI systems to understand the world in a meaningful, reliable, and explainable manner. By creating these internal representations, AI systems can effectively interact with the world, navigate intricate environments, engage with others, and make decisions that rely on observation and beliefs.

The grounding of AI is the goal of Knowledge Representation and Reasoning (KRR), a foundational aspect of AI research that focuses on how knowledge about the world can be effectively represented, organized, and understood by AI systems. The goal of KRR is to enable AI systems to understand and reason about the complex and dynamic environment in which they operate. The task of KRR is to translate the human-created model of the world for AI to reason about and improve.

This is important for AI systems, as grounding bridges the gap between them and humans, providing a shared basis for meaning, context, and understanding. It enables autonomous agents to navigate complex environments, interact with others, and make decisions that respect societal norms and ethical principles.

Overcoming the problem of grounding involves creating AI systems that can meaningfully link their internal representations, which are often abstract and detached from reality to the actual objects and phenomena they represent. In other words, these AI systems must understand causal relationships and causal reasoning.

Grounding would ensure that city-scale autonomic systems work seamlessly with human operators, adapt to dynamic urban environments, and prioritize citizens' well-being and safety. For planet-scale autonomic systems, grounding facilitates effective collaboration between AI systems from different countries, cultures, and backgrounds, aligning the global impact of AI technologies with human rights and shared values.

> To ensure that we can effectively steer AI systems as they become increasingly autonomous, grounding them early on will be crucial.

Solving the grounding problem is urgent and fundamental to the regulation of AI, autonomous agents, and the future autonomic computing network (Web 3.0) because it addresses the core issue of aligning AI systems by encoding in the algorithms human values, intentions, and understanding. This requires standards that take into account more than purely technical aspects.

4.5 Socio-technical Standards Bridge the Gap

Key Takeaways:

- Socio-technical standards propose a regulatory approach for AI and AIS in the Web 3.0 era, integrating technical, social, legal, and ethical aspects.
- These standards prioritize adaptability, collaboration, and stakeholder participation, ensuring compliance, safety, and alignment with societal values.
- Examples include aerial drone regulations and self-driving car standards that address technical, legal, and social factors.
- Global socio-technical standards are crucial for technology to respect human values, protect rights, and promote collective well-being.

Socio-technical standards are unique because they aim to integrate and balance the technical, social, physical, and legal use of technology. They are not merely technical solutions; rather, they are technical solutions that are expressly designed to meet social, legal, and ethical requirements by enabling machine-readable, machine-shareable, and machine-executable "laws as code."

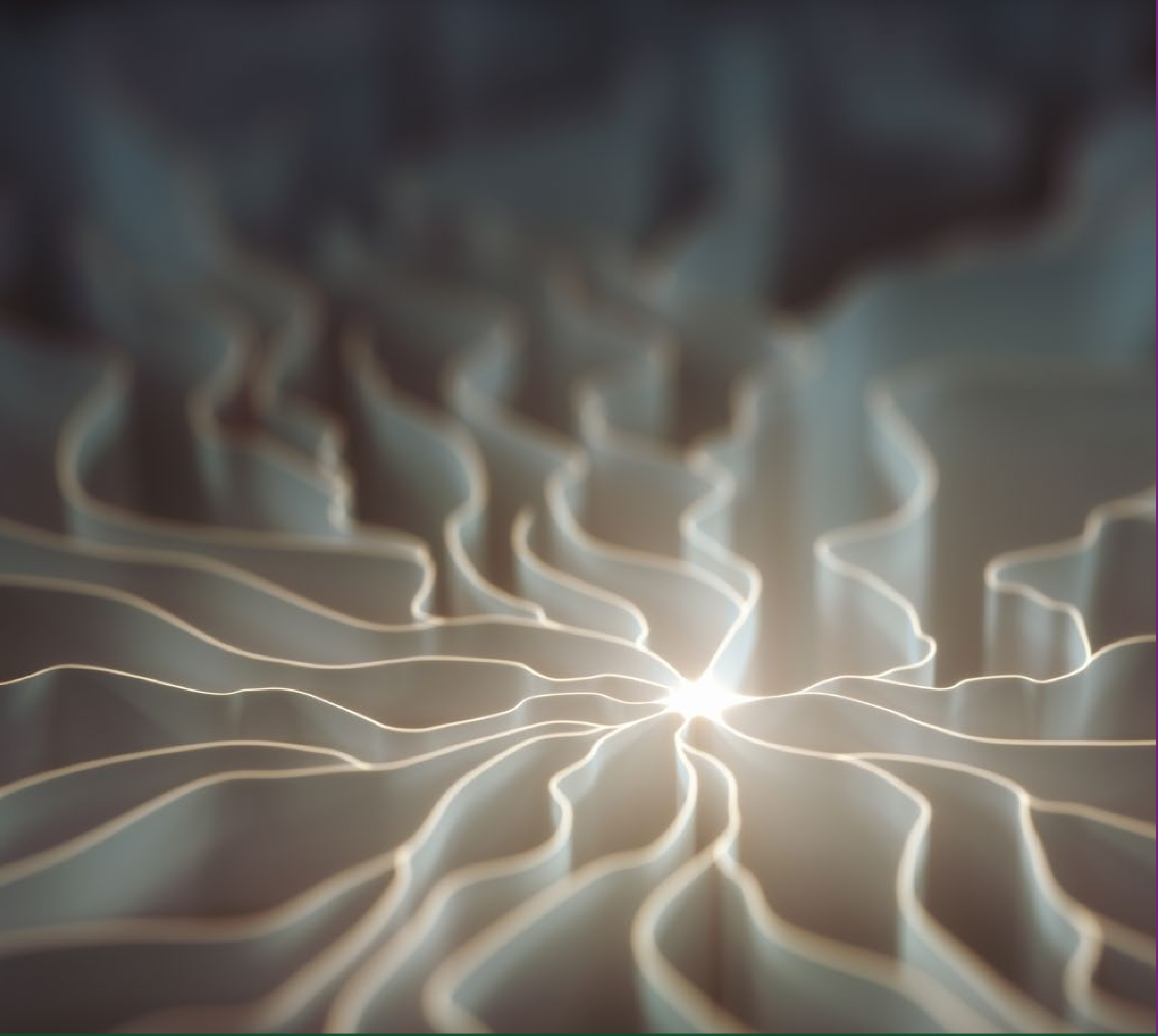
Because of the complex nature of AIS, which interact in both digital and physical realms, through software and hardware (like IoT devices, robots, and drones), socio-technical standards are an ideal governance solution. These standards would incorporate pertinent laws, rules, and regulations as a baseline (or "bare minimum") level of management, while also incorporating ethical considerations by drawing upon interdisciplinary knowledge. Ultimately, their design would allow for the continuous evaluation and improvement of AI systems, ensuring effectiveness and relevance while fostering safe and harmonious interactions that align with human values, now and into the future.

For example, socio-technical standards applied to AI-operated aerial drones would encompass both technical, spatial, legal, and social aspects. They address technical considerations like flight capabilities, data security, and autonomous features, alongside social factors such as privacy concerns, laws and safety regulations, and ethical considerations.

To illustrate, imagine what a socio-technical standard governing drones might look like. The standard would first integrate basic laws, such as a requirement for the drone to adhere to a posted speed limit of no more than 20 miles per hour. The drone would be programmed to avoid flying over schools, government buildings, or spaces with a maximum population density threshold, or other areas where privacy and safety are paramount, even if this is not specifically required by law. The standard could also require the drone to apply facial blurring algorithms when capturing footage of crowds. By not only complying with basic laws, but also promoting responsible and ethical drone usage, socio-technical standards could therefore safeguard privacy, enhance safety, and foster increased public acceptance of drones in applications like infrastructure maintenance, deliveries, and human transportation.

As a second example, consider a socio-technical standard for self-driving cars. The standard first incorporates the local vehicle laws that require each AV to do basic things that all cars with human "drivers" are expected to do, such as obey the posted speed limit, or come to a complete stop at a stop sign. The standard could also integrate certain "best practices" typically encouraged by traffic schools—such as avoiding tailgating by allowing at least two car-lengths between vehicles—even if these are not always adopted by human drivers. The resultant benefits could take the form of better traffic flow, lower energy consumption, or drastic reductions in the number of fatal car accidents.

The development of global socio-technical standards therefore represents a crucial step in our collective journey toward a future where technology enables governments and markets to harmonize with diverse human values and perspectives. Such standards could ensure that powerful and transformative technologies are developed and used in ways that respect our societal values, protect our legal rights, and promote our collective well-being, while still achieving technical excellence and innovation. These standards would, if adopted, represent an important step forward in our journey toward a future where technology serves humanity, not the other way around.



PART 5

Socio-technical Standards for the Spatial Web

5.1 A Public Imperative

Key Takeaways:

- Early incorporation of socio-technical standards is crucial for aligning AIS with human values, intentions, and understanding, reducing the risk of harmful behavior.
- The IEEE launched the Global Initiative on Ethics of Autonomous and Intelligent Systems in 2016 to address ethical, legal, and social concerns in the development and deployment of AI and autonomous technologies.
- The IEEE P2874 Working Group is developing socio-technical standards to support the alignment, interoperability, and governance of AI and AIS.
- Socio-technical standards ensure that emerging technologies are both technically sound and socially beneficial, respecting societal values and promoting collective well-being.
- These socio-technical standards may serve as the foundation for the AI-powered Web 3.0 and offer a method to steer society toward an autonomous future.

In April 2016, the IEEE—the world's largest technical professional organization with more than 420,000 global members involved in all technical areas pertaining to computer science, electronics, and electrical engineering—launched the Global Initiative on Ethics of Autonomous and Intelligent Systems.

Later that year, the IEEE P7000 series was initiated to address the ethical, legal, and social concerns that arise from the development and deployment of AI and autonomous technologies, with the aim of ensuring that these systems are designed and used in ways that align with human values and societal norms. The P7000 series covers a wide range of topics, including transparency, accountability, data privacy, and algorithmic bias.

The output of the P7000 series, a paper about Ethically Aligned Design (EAD), "A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," identified issues and provided recommendations in key areas pertaining to AIS. It provides a framework for developers, manufacturers, and regulators to ensure that AI technologies are ethically designed, implemented, and governed.

Historically, there have been two broad categories of standards: technical standards and social standards. With the rise of AI and its growing impact on various aspects of our lives, including personal privacy, public infrastructure, cultural norms, and the economy, there is a clear requirement for a new type of standard: socio-technical standards. Designed to bridge the gap between technical and social domains, these standards ensure that emerging technologies are not only technically sound but also socially beneficial and aligned with societal norms and values.

While the concept of socio-technical standards is not new, its application to AI is a recent development. By implementing socio-technical standards for AIS, powerful and transformative technologies gain the potential to expand in ways that respect our societal values, protect our legal rights, and promote our collective well-being, while still achieving technical excellence and innovation.

In 2020, the Spatial Web Foundation (SWF), a non-profit SDO (and co-author of this paper) whose mission is to develop socio-technical standards, partnered with the IEEE to lead the development of state-of-the-art socio-technical standards and protocols. The goal of these standards is to define the standardization and guidelines for the implementation of AIS in alignment with the IEEE P7000 guidelines. **The IEEE has declared this effort a "public imperative."**

The Spatial Web Foundation presented the first draft specifications at the launch of The Spatial Web Protocol, Architecture and Governance Working Group (IEEE P2874 Standard) on July 21st, 2021. This group consists of members from academia, government, and various industries. The launch of this working group marks the public commitment toward the development of a 21st-century "cyber-physical" Spatial Web.

The Spatial Web standards are both social and technical, designed to support the alignment, interoperability, and governance of AI and AIS. According to the Spatial Web Foundation, the standards and protocols, when adopted and widely implemented, will form the foundation upon which the AI-powered Web 3.0 is built. Following, we highlight a few essential components of the proposed standards.[1]

5.2 HSML: Modeling Human Knowledge and Experience

Key Takeaways:

- HSML (Hyper-Spatial Modeling Language) is a knowledge modeling language that enables systems to encode properties of physical objects, logical concepts, and contextual activities linking people, places, things and AI. HSML facilitates multimodal modeling and knowledge sharing among machines and humans, encompassing ethical, moral, economic, and societal considerations.

Hyper-Spatial Modeling Language (HSML) is a knowledge modeling language that is designed to encode the properties of any physical object or concept, as well as the context of any activity, in digital or physical space. The purpose of this knowledge modeling language is to enable multimodal (multisensory) modeling, converting data from text, audio, video, 3D, or other sensor-based datasets, allowing machines to model, interpret, and share knowledge with one another and with humans. In effect, HSML is a common language that "grounds" semantic, sensory, and social information into a common data representation or "shared world model."

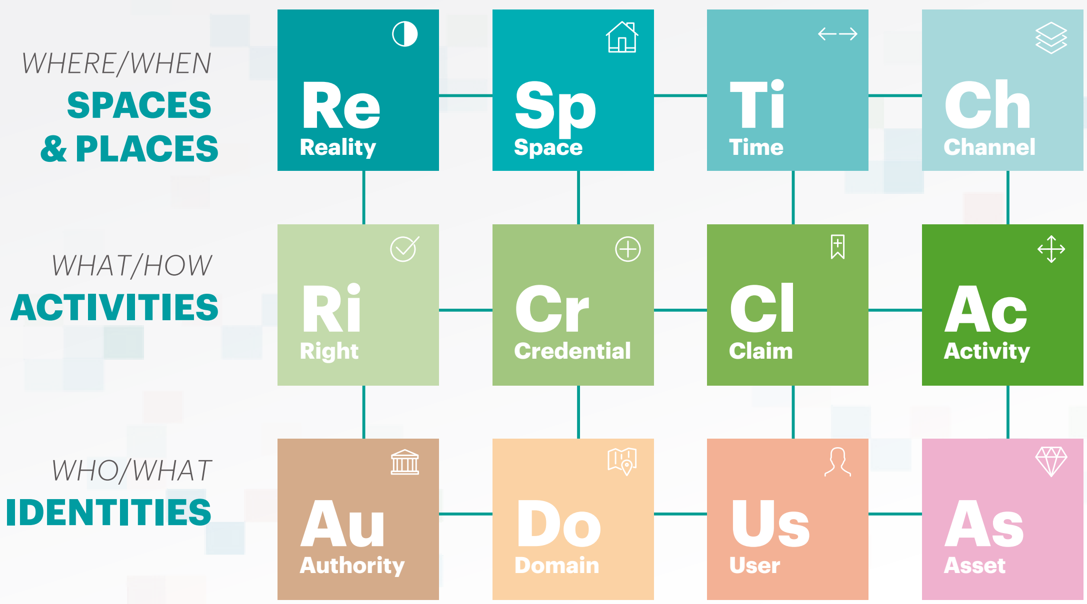
HSML is structured to allow for the creation of rich data models using a set of contextual elements [see the chart on the following page] that can represent a wide range of entities, policies, identities, places, and activities.

With HSML, machines can update their understanding of the human domain and therefore interact with us and the world more effectively, while humans can interpret what machines understand. **Through a shared world model, machines are able to comprehend the causal factors, circumstances, timing, underlying mechanisms, and principles governing any given activity, while humans have a transparent and explainable mechanism for interpreting AI decision making. The model also enables the expression of ethical, moral, economic, and societal perspectives.**

Such a model could equip AI with the ability to comprehend multiple perspectives simultaneously, resulting in a richer, more nuanced, and more comprehensive understanding of human needs and activities.

1 This summary of the proposed Spatial Web Standards is meant to provide an overview of the standards being proposed by the Spatial Web Foundation through its work with the IEEE. This summary of the Spatial Web Standards is not intended to reflect an endorsement of these standards by any individual author of this paper, including Dentons US LLP.

HSML Contextual Elements

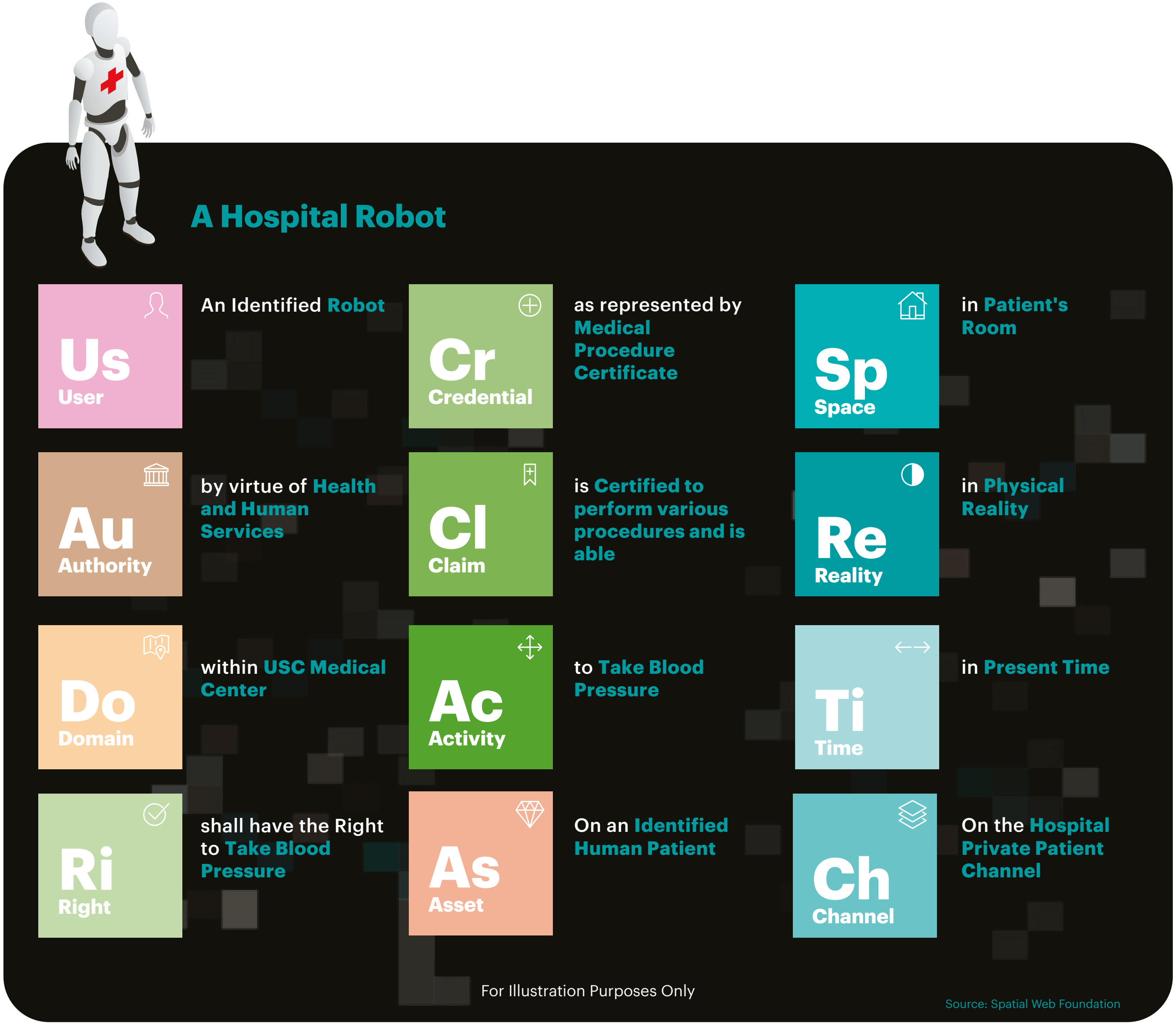


For Illustration Purposes Only

Source: Spatial Web Foundation

HSML supports the creation of custom schemas that enable a multidimensional description of the context. Context is understood as the combination of the elements and of the relationships between entities, objects, locations, and actions—commonly known as the Who, What, When, Where, How, and Why of any scenario, situation, or circumstance. The answers to these questions are often stored in different data silos and different data spaces. They need to be made interoperable, shareable, and addressable by multiple competing AI algorithms that can maintain their coherence at scale.

In the chart below is an example of how HSML parses natural language into elements that enable a hospital robot to understand and interact with us and the world safely, securely, and more effectively.



Because of the ability to parse natural language and sensory data into code and therefore understand context and causal relationships, HSML is uniquely positioned to meet the requirements of "grounding" across multiple perspectives, ensuring that AI systems can effectively interpret, understand, and share knowledge with one another and with humans in a diverse range of contexts.

5.3 Integrating Perspectives

Key Takeaways:

- By incorporating subjective, objective, intersubjective, and interobjective grounding, AI systems can comprehend and align with human perspectives, enabling effective communication, collaboration, and ethical deployment. This comprehensive approach ensures adaptability, context awareness, and ethical alignment of AI technologies with human values and societal norms.

Because of the ability to parse natural language and sensory data into code and therefore understand context and causal relationships, HSML is uniquely positioned to meet the requirements of "grounding" across multiple perspectives, ensuring that AI systems can effectively interpret, understand, and share knowledge with one another and with humans in a diverse range of contexts. HSML enables machine modeling of multiple human perspectives, which can provide a holistic understanding of different aspects of reality and meaning for AI systems.

1. **Experiential Grounding (Subjective Perspective):** This grounding type focuses on modeling, understanding, and reasoning about symbols and their meanings (semiotics) and their relationship to real-world entities or concepts from the perspective of human subjective experience. This understanding is crucial for AI systems to comprehend natural language inputs and generate appropriate responses that align with our individual perspectives.

2. **Physical Grounding (Objective Perspective):** This grounding type emphasizes the importance of modeling physical phenomena and interactions with the environment and the context in which interactions occur. It reflects the objective aspects of individual sensory and environmental experiences, enabling AI systems to develop an embodied understanding of the world and its interactions with it via sensors across the IoT.

3. **Cultural Grounding (Intersubjective Perspective):** This grounding type concentrates on the role of modeling the intersubjectivity of human social dynamics, considering shared experiences and social conventions, cultural norms, ethics, and values. By modeling the collective, intersubjective human experience, AI systems can engage in empathetic and emotionally intelligent interactions with humans and other agents, aligning with our social and emotional perspectives.

4. **Structural Grounding (Interobjective Perspective):** This grounding type focuses on modeling interobjective knowledge, real-world entities, or categories and their relationships within the larger collective systems humans operate in, such as social, technological, legal, economic, and ecological systems. By modeling the interobjective aspects of human knowledge, AI systems can better comprehend the complexities and dynamics of the structure of the world in which we operate.

By incorporating these four major human perspective categories into the development of AI systems, we can take a more integral approach to enabling adaptable, ethically aligned, and contextually aware AI agents. This comprehensive approach to grounding ensures that AI technologies can successfully interact with humans, other AI systems, and their environments, while considering the diverse perspectives and complexities present in any given situation from a human point of view.

Adopting HSML would allow AI systems to explicitly model (and contribute to) a more comprehensive understanding of human perspectives across various contexts.

Source Spatial Web Foundation

5.4 HSTP: Authentication and Communication for Data and Devices

Key Takeaways:

- HSTP (Hyper-Spatial Transaction Protocol) provides the methods for passing HSML messages in the Spatial Web. It provides a universal, secure, and verifiable protocol for communication between digital or physical systems, ensuring seamless interaction and cooperation between diverse AI systems. It incorporates a zero-trust architecture and strict authentication measures for secure data exchange and control over AI operations.

The Hyper-Spatial Transaction Protocol (HSTP) is an innovative technology that bridges the gap between digital and physical realms, enabling the enforcement of AI activities across both domains. By providing a universal, open protocol for transferring digital or digitized physical content, HSTP creates a common language that all systems can understand, thereby facilitating seamless message passing, interaction, and cooperation between diverse AI systems.

HSTP's zero-trust architecture, which mandates verifiable credentials for any interaction between systems, ensures data exchanges across environments is done so with security at top-of-mind. This rigorous credential-based authentication process is particularly crucial for AI activities, as it allows them to operate within a secure and reliable framework, protecting against unauthorized access, ensuring the integrity of data and operations, and significantly enhancing the security and trustworthiness of all operations across the Spatial Web.

In contrast to the open structure of the Internet and the World Wide Web, the Spatial Web, based on HSTP, is designed as a permissioned network. This fundamental shift in architecture not only enhances security but also increases the reliability and predictability of AI operations within this environment. Each entity on the Spatial Web—be it a user, device, or AI—is assigned a unique Spatial Web ID (or SWID). This ID would serve as an authenticated digital identity, enabling precise control over access permissions and ensuring that each entity can only perform the actions it is authorized to carry out. Spatial Web IDs can be federated between themselves and/or be federated with other digital ID systems depending on the context.

The use of verifiable and authenticated Spatial Web IDs adds an extra layer of protection against various cybersecurity threats. It guards against impersonation attacks by verifying the identity of each participant in an operation, preventing unauthorized access and manipulation of data. It also protects against data breaches by ensuring that only authorized entities can access sensitive information.

In addition, HSTP's stringent authentication measures can mitigate cyber-physical risks. In a world increasingly populated by AVs, drones, robots, and other AI-driven systems, the ability to authenticate and control in a reliable manner these entities' actions is crucial. By requiring verifiable credentials for each operation, HSTP can prevent unauthorized control of these systems, protecting against potential threats to public safety and infrastructure.

Finally, by enabling seamless and secure data transfer across real and virtual spaces, HSTP allows for the integration of AI activities with the Spatial Web. This fusion results in an enhanced level of collaboration between humans and machines, with AI systems capable of operating and interacting within the same shared digital and physical contexts as humans.

Through HSTP, AI-powered activities can be conducted across a wide range of applications, from autonomous vehicles navigating in real-world traffic to AI-driven logistics optimizing goods transportation in a smart city to virtual assistants interacting with users in augmented reality environments. The result is an interconnected intelligent network that spans the globe, creating a safer, more efficient, more collaborative, and ultimately *smarter* world.

PART 6

Building A Smarter World

6.1 A Formula for Adaptive AI Governance

Key Takeaways:

- The challenge is to create policies and regulations that can be understood and applied by both humans and machines, bridging the gap between traditional human-centric drafting and machine-readable and executable regulations.
- Translating existing regulations into machine-interpretable code allows for the governance of machine behavior in compliance with policies, along with the ability to adapt to regulations as they change over time.
- The use of Spatial Web standards could enable automated interpretation, enforcement, and dynamic compliance with laws, while also identifying and addressing ambiguities in natural language laws to enhance their machine-readability.

Throughout human history, nearly every law, rule, regulation, and court decision has been based on the assumption that the subject being governed is a *person, either a human being or human led organization*. The current legislative and regulatory frameworks in place were primarily designed by humans, and for humans. However, with the emergence of human-level AI, this may no longer hold true, presenting an unprecedented challenge: **How can policies and regulations be crafted in a manner that is understandable and applicable to both machines and humans?**

To answer this question, we would need laws that are machine-readable, machine-interpretable, and machine-executable. This entails the translation of legislation, court rulings, regulatory guidelines, etc., into formats that can be interpreted by and applied directly to intelligent non-human entities.

Take, for example, driving laws in various US states, and globally. Many of these laws define a "driver" as a "person" in "physical control of a vehicle." In the case of self-driving cars, where a human is often not physically controlling the vehicle, these definitions will need to be re-examined.

Translating existing laws, policies, and regulations into computer code is sometimes referred to as "rules as code" or "law as code." Expressing laws in machine-readable code enables machine behavior to be governed directly. Whether supervised by human beings or operating autonomously, machines adhere to the appropriate policies. This approach also offers the flexibility to dynamically adjust laws and rules over time as needs change. For example, speed limits could be adjusted in real-time to factor in inclement weather. However, the challenge lies in determining where technology can replace human interpretation and where it cannot. The challenge of "rules as code" also becomes complicated when varying legal systems conflict with one another (e.g., differences between common and civil law jurisdictions).

In addition to machine-readable code, AI-enabled machines must possess the technical dexterity and acumen to rapidly switch and adapt to contextual variations, which depend on their operational environment. Since regulations may be written in different languages, and for different use cases, achieving scalable machine interpretability and machine executability requires the development of a semantic representation, i.e., of a format for law as code that can capture the meaning of words, sentences, or concepts used in natural language law, in a structured and meaningful manner. This representation must be able to accommodate regulatory changes and multilingual regulations across different regions, in real time.

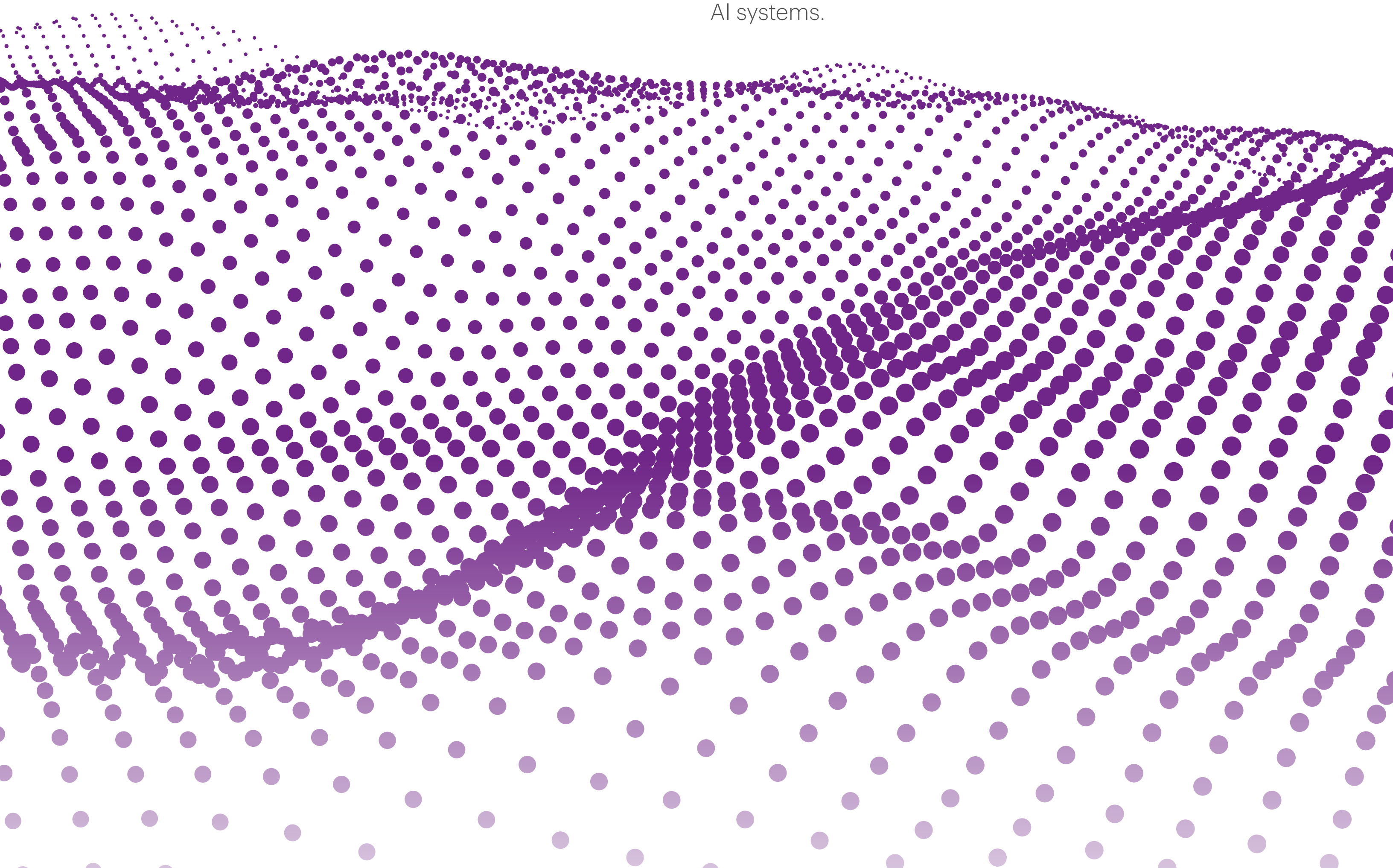
One potential avenue for addressing these challenges lies in the adoption of Spatial Web standards. These standards offer a comprehensive framework for modeling rights, permissions, relationships, and contextual information regarding actions, locations, timing, objectives, duration, and the involved parties or entities. In essence, they enable the modeling of various types of conduct, principles, and rules that govern the interactions between humans and machines in a way that transcends language barriers and textual limitations.

Spatial Web standards also enable users and stakeholders to encode the properties of physical objects and concepts, including contextual information, coordinates, and credentials. This provides any AI system that adopts the Spatial Web standards with a comprehensive representation of the relationships between entities, policies, identities, and activities within their world models. By leveraging these standards, AI systems can develop semantic representations applicable to specific geolocations that are both machine and human friendly.

> The ability to comprehend and interpret region-specific regulations and cultural subtleties enables Spatial Web standards to cater to the specific needs of different regions while also maintaining global interoperability.

In essence, these standards are designed to be interoperable while remaining adaptable to local contexts.

The technological challenge of "adaptive" AI governance is twofold. Beyond explicit world models, semantic representations, and differing legal jurisdictions, AI and AIS must possess the capability to dynamically adapt to changing environments and contexts while maintaining reliable and trustworthy outcomes. This may require exploring new approaches to AI that can reason causally, going beyond today's machine learning-based generative AI systems.



FF2020: A Real-World Application of Socio-technical Standards

The European Union's Flying Forward 2020 (FF2020) program provides real-world use cases that demonstrate how Spatial Web Standards are working toward solving the challenges of law as code. By translating existing legal requirements, decisions, and regulations into HSML, the behavior of UAVs can be governed in accordance with legal policies.

The FF2020 project addressed these challenges by developing a geospatial infrastructure using draft IEEE Spatial Web standards & protocols (HSML and HSTP) to systematically enforce rules and policies within the domain of Urban Air Mobility.

Through machine-readable models and data integration from various sources, FF2020 enables autonomous drones to comprehend and seamlessly comply with laws and conditions. The project has successfully translated existing EU laws for Urban Air Mobility into Hyper-Spatial Modeling Language (HSML), allowing autonomous drones to adhere to parsed rules and regulations automatically, even those that inherently contain ambiguities. For example: "UA operators must fly safely." But how is "safely" defined? HSML can explicitly identify ambiguous laws and provide a feedback mechanism for developing future analog laws with machine-readability in mind. This creates a virtuous two-way loop between analog law and digital law.

FF2020 has demonstrated its effectiveness through real-world use cases, including emergency delivery, infrastructure maintenance, and security monitoring. The project's approach ensures compliance with local rules and regulations without manual intervention or human involvement. It also enables autonomous drones to navigate airspace, avoid security and no-fly zones, and understand and enforce regulations set by the European Union Aviation Safety Agency (EASA).

One example is a rule in 'UAS.STS-02.010 General provisions' on the procedures for the operation of unmanned aircraft. When flying an unmanned aircraft within a horizontal distance of 50 meters from an artificial obstacle taller than 105 meters, the maximum height of the UAS operation may be increased up to 15 meters above the height of the obstacle.

FF2020's approach has the potential to revolutionize policy and law creation, sharing, and implementation worldwide. By embracing digital accessibility, machine-readability, and automation, it enhances service delivery, clarity, consistency in laws, and expedites decision-making. The goal is to have legislation coded and managed by relevant organizations and governments, optimizing rules for autonomous systems and improving comprehension and execution for civil servants and citizens.

By embracing this approach, governments can tackle the complexities of rule creation, sharing, and interpretation. Ambiguous rules, when coded into software, can introduce inconsistencies in accessible benefits and services for citizens. Additionally, software developers often struggle with implementing rules designed for outdated technology, lacking the ability to influence rule development. Aligning law with software and code is crucial for building a legal industry compatible with our software-driven world.

6.2 Active Inference: Self-Regulating Intelligence

Key Takeaways:

- Active Inference-based AI operates on world models, understanding causal relationships and enabling continuous learning and adaptation.
- It can facilitate the explainability of AI systems by explicitly modeling their decision-making processes.
- Using shared world models, it can meet regulatory requirements while creating AI systems rooted in biological intelligence for a prosperous, sustainable, and equitable future.
- This approach allows the creation of a distributed ecosystem of AI agents, forming a multi-dimensional and cyber-physical web that connects physical and virtual spaces; realizing the Spatial Web.
- It enables networks of small, agile, and adaptive AI agents that can share knowledge, explain their actions, ask questions, and work together with humans to solve higher-order complex tasks, leading to a reshaping of how laws are made and enforced.

The future of AI must not be determined by "black box" large language models or a race to a monolithic planet-scale super intelligence operated by a single company, organization, or government. Instead, AI could take a more natural path to general and even super intelligence in the form of smaller, agile, adaptive and autonomous "Intelligent Agents" that share knowledge, explain their actions, ask questions, and that are even "curious" about the world. Unlike today's complex, data-hungry AI systems, these agents would require minimal training and would rely on smaller amounts of highly contextualized data—what one might call "smart data," as opposed to the currently ubiquitous "big data" approaches. They would be specialized to accomplish specific tasks at an expert level. However, they would also have the ability to communicate with one another, continuously exchanging knowledge in order to tackle complex challenges, all while retaining the capacity to acquire new information. Essentially, these agents would operate in a manner akin to human intelligence.

The human brain is a remarkably complex organ capable of learning, planning, and making decisions more efficiently and elegantly than any state-of-the-art AI system. By understanding the brain's inner workings and applying them to the field of computing, we may be able to develop Autonomous Intelligent Systems that are generally and genuinely more intelligent.

Karl Friston, renowned British neuroscientist, theoretician at University College London, and Chief Science Officer at VERSES, is one of the world's leading authorities on brain imaging and theoretical neuroscience. Friston specializes in the use of physics-inspired statistical methods to model neuroimaging data, brain mapping, and other dynamic systems, including complex living systems. Ranked #1 by Semantic Scholar in the list of most influential neuroscientists, Friston has made notable contributions to our understanding of human cognition and social behavior through his theory of intelligence known as "Active Inference."

Active Inference is a way of understanding cognitive functions and sentient behavior in living systems. The framework has demonstrated its applicability in explaining and modeling a wide range of phenomena, from the activities of bacteria to the evolution of the brain. At its core, Active Inference aims to minimize "free energy," which represents the surprise or discrepancy between expected and actual sensory information.

Developed through groundbreaking research in neurology and physics, Active Inference utilizes a statistical method to model the systems responsible for decision-making, perception, and action. Additionally, it describes perception, planning, and action using probabilistic inference, a mathematical manner of determining how likely some outcome or event is to happen.

This widely accepted theory has already been applied across various disciplines, including neuroscience, biology, psychology, philosophy of mind, the social sciences, and robotics. More recently, it has found applicability in the realm of AI. In essence, Active Inference provides the mathematical blueprint for constructing world models that intelligent agents can use for planning and taking action.

Researchers from VERSES (co-authors of this paper), the Wellcome Centre for Human Neuroimaging at University College London, the Departments of Cognitive Computing and Philosophy at the Université du Québec à Montréal, and the Berlin School of Mind & Brain at Humboldt-Universität zu Berlin, published a landmark research paper in June 2023 on designing explainable artificial intelligence using active inference.

The paper describes methods for developing explainable, auditable AI systems, and proposes an architecture for decision-making in AIS that are equipped with human-like introspection abilities. More specifically, it formalizes the techniques that model how AI systems make decisions and select appropriate actions based on incomplete information (or "uncertainty").

Active Inference utilizes two crucial capabilities: "self-modeling," involving the creation of internal representations, and "self-access," allowing AI systems to analyze their internal states and decision-making processes. These capabilities facilitate a deeper understanding of an AI system's decision-making mechanisms and enable the system to provide self-reports. This self-modeling and self-access is fundamental to self-improvement and learning. By continuously assessing themselves, Active Inference-enabled AI systems consistently improve their outcomes and can at the same time provide detailed explanations for their actions.

While this explainability is crucial for building trust and understanding among end users, Active Inference has an even more substantial benefit: it could lead AIS to become truly autonomic. This is because Active Inference-based AI agents naturally seek equilibrium within the ecosystems where they operate. In striving to maintain a healthy, balanced role, they operate and adapt in the most optimal manner given the situational context.

Their continual pursuit of systemic equilibrium is possible because, unlike generative AI systems that predominantly seek patterns in data, Active Inference-based AI systems model the causal factors that generate the data. They operate on explicit and explainable world models, which represent their "understanding" of their environment and the underlying causal relationships between data points and the activities that cause them. Consequently, Active Inference agents can self-regulate by learning, adapting, and continuously updating their world model in real-time to align or seek out "fitness" with environmental changes.

Such agents can accomplish goals by prioritizing a path of least resistance, attempting to find the best outcome with the least amount of energy. This method, based on the "principle of least action" from physics, defines intelligence as the process of determining and achieving the optimal outcomes, given limited data (uncertainty)—leading to agents that are not only more intelligent, but also maximally efficient, requiring the least amount of energy to achieve their goals. As such, one can say that

> Active Inference provides us with a mathematical principle for designing autonomous intelligent systems, based on the physics of intelligence as it occurs in nature.

Smarter AI governance could be achieved by focusing on intelligent agents that exhibit three crucial characteristics: explainability, self-improvement, and the ability to maintain dynamic equilibrium (i.e., resilience and sustainability).

These traits offer significant advantages that align with growing demand from governments, regulators, policymakers, and human interest groups for AI systems to be more interpretable and auditable by users and stakeholders.

By combining this class of agent with the Spatial Web standards (HSML and HSTP), a new class of *autonomic* intelligent agents becomes possible. The potential is awe-inspiring, and represents, as the authoritative book on the Spatial Web suggests, a new Intelligence Age in which humans and machines work together to build a smarter, safer, better world for us all.

Aligned AIS at Work: AI in Earthquake Disaster Management

Let's consider a situation where this kind of next-generation Active Inference-enabled AI agent is leveraged as an integral part of disaster management during an earthquake event. The agent has to understand human emotional states, coordinate with various agencies, provide real-time updates, and prioritize resources based on available and changing data.

In the immediate aftermath of a devastating earthquake, the agent, operating on socio-technical standards, and monitoring local systems, springs into action. It begins by forming a set of initial assumptions or "priors" about the state of the world into a world model, based on the first influx of information it receives. This information comes from social media, phone calls, and emergency alerts, as people make urgent requests for help. The agent's understanding of language and semiotics helps it comprehend these human distress signals, allowing it to generate initial plans on where immediate help is required.

Simultaneously, the agent taps into a network of seismographic sensors, drones, and cameras to collect crucial physical data about the quake. It uses this data to update its world model, refining its assumptions about which areas are likely to be hit by aftershocks, the extent of damage, and routes blocked by debris. This sensory and environmental understanding informs its decision-making, such as directing a fleet of drones to clear specific paths for rescue vehicles or deciding to airdrop emergency supplies where they're most needed.

As the agent processes the enormity of the crisis, it continually updates its model of the collective sentiment of the affected population. It recognizes changes in public concerns, fears, and needs expressed on social platforms and news channels. This "understanding" of shared human experiences helps it dynamically adapt its plans for resource distribution. For instance, if it anticipates an unexpected increase in distress over lack of water in a specific area, the agent can confirm this via other agents operating local drones, cameras, and sensors and respond accordingly by reprioritizing its resources to set up water purification systems in that area.

Finally, the agent iteratively refines its "understanding" of the various organizations involved in disaster response. It learns to coordinate with NGOs, government agencies, healthcare services, and more, matching their capabilities and proximity to the affected areas with the requirements on the ground. This representation of the structure of collective systems allows the agent to make the most of the resources at hand. For example, it might adapt its strategy to guide medical personnel and robot assistants from a nearby hospital to an area where severe injuries have been reported.

The power of this approach relies on the combination of Active Inference and socio-technical standards that enable the complex modeling of human knowledge and experience. These systems effectively model our world, learn from unfolding situations, adjust their expectations, align their predictions with reality, modify their actions, and collaborate with humans and other AI agents to construct strategies. Agents understand and interpret situations from multiple perspectives, providing them with a holistic view of the situation. The integration of data and understanding allows it to operate at a level that is both highly informed and highly adaptable, leading to more efficient, effective, and empathetic disaster response.

Imagine similar agents and AIS, coordinating city traffic, optimizing healthcare, overseeing industrial production lines and optimizing supply chains. Whether it's managing energy grids, tailoring education, supporting elder care, enhancing cybersecurity, or mitigating disasters, these agents demonstrate an elevated level of understanding, responsiveness, and adaptability that opens up new solution possibilities. This standards-based approach could transform numerous scenarios, big and small, ushering in a new era of adaptive distributed intelligence at scale.

Conclusion: Smarter Governance in the Intelligence Age

As we enter the Intelligence Age, our concepts of governance and its function in society must adapt and become smarter. The immense potential of Autonomous Intelligent Systems requires new regulatory frameworks that harness AI's capabilities while effectively mitigating its risks. Traditional legislative approaches must therefore evolve to include not only the governance of individual AI systems but of networks of AI systems. Crucially, this evolution should consider that the governance of intelligent machines may ultimately take the form of sophisticated, dynamic, and collective self-governance.

Smarter governance is possible through the application of socio-technical standards and smarter approaches to AI such as Active Inference. Socio-technical standards can be the guiding force that steers AI toward its greatest gifts, empowering us to develop increasingly autonomous systems that are explainable, interoperable, governable, and thus safe, fair, and aligned with our values.

The shared understanding between humans and AI that socio-technical standards facilitate can pave the way to global interoperability, while simultaneously ensuring compliance with regulatory requirements, cultural norms, and ethical considerations on a regional scale. Socio-technical standards allow AI systems to become trustworthy, to network, and to one day even govern themselves, while also granting us the ability to define the desired level of human oversight and involvement.

By adopting socio-technical standards, we have the opportunity to safely weave our digital and physical realms into an AI-powered Spatial Web, enabling an unprecedented level of communication between humans and AIs, and between AI systems themselves.

Socio-technical standards can enable a new class of self-learning, adaptive and self-regulating AI and AIS. These systems, based in biophysics and neuroscience, could unlock dynamic governance systems that will be required to effectively regulate them. Smarter agents could serve as the catalyst for this transformation, driving effective decision-making and paving the way for a smarter governance.

Unlike the standards and protocols of the World Wide Web, where privacy and security were afterthoughts, the Spatial Web standards outlined in this paper contain universal, secure, and verifiable communication protocols for authenticating interactions. This architecture significantly reduces security risks associated with AI systems by granting access to data and devices to explicitly authorized and approved entities, with the ability to revoke or restrict their access or activities at any time.

Spatial Web standards could become the foundational modeling language and communication protocols through which humans collaborate with intelligent machines. These standards are the first and perhaps most essential step toward building autonomous and autonomic networks. The outcome could be a planet-scale network of AIS like the Spatial Web. This web of everything will not only power smart devices, smart cars, and even smart cities—but an actual "Smarter World."

In this "Smarter World" world, AI adapts seamlessly to real-time changes, enhancing our safety and enjoyment. Similar to how a human functions without constantly worrying about the operations of their respiratory or digestive systems, AIS based on Active Inference can determine on our behalf the optimal management of tasks. In a Smarter World, AI responds to humanitarian and environmental crises in ways that surpass the limitations of human ingenuity. AI enables safer and more efficient transportation, more effective and affordable healthcare, more personalized and accessible education, and a global supply chain that operates with unprecedented levels of resilience and optimization, reducing waste and energy consumption.

With AIS based on socio-technical standards, we have the power to make a Smarter World a reality.

The future of AI governance, therefore, is not merely about restrictions and controls designed to avoid the potential perils of AI. It should also provide us with enormous benefits by steering AI systems toward their most promising capabilities. By liberating us from menial, repetitive, exhausting physical and mental labor, properly governed AI systems will allow people more bandwidth to focus on the critical issues that we face as individuals and as a civilization. Smarter AI governance aligns AIS with our most vital human values while empowering these systems to operate more autonomously, more intelligently, and more harmoniously at global scale.

We can begin our journey toward smarter AI governance now by testing innovative products, technologies, laws, and business models related to AI and AIS using interoperable socio-technical standards (see annexe on The Prometheus Project).

The true prize of the Intelligence Age is the freedom to augment, enhance, and extend our genius; to focus on creativity, innovation, compassion, and individual and collective self-actualization, pushing the boundaries of what is possible for ourselves and for future generations.

> To that end the future of AI governance is not only about steering AI, but about how we can use technology to steer ourselves down a brighter path toward the boundless potential of human imagination.

ANNEXE: The Prometheus Project

In an era marked by escalating geopolitical tensions, establishing a universally recognized socio-technical standard for autonomous systems could serve as a unifying catalyst. By fostering seamless communication and interaction between different autonomous systems across nations, socio-technical standards can help avoid potential conflicts arising from differing technological approaches and political ideologies.

In a world where global challenges are increasingly complex, utilizing AI to work collaboratively across borders and political boundaries is crucial. Regardless of the country of origin or technological approach, autonomous systems should be able to work together to solve complex problems, improving efficiency and productivity in various industries and territories worldwide.

Twenty-first-century challenges call for twenty-first-century solutions. As such, the Spatial Web Foundation (SWF) is calling for the creation of a global regulatory simulation sandbox utilizing state-of-the-art AI and advanced simulation technology that runs on the Spatial Web standards.

The Prometheus Project, named after the Titan God of Fire, is an invitation to form a public-private initiative aimed at testing and exploring the potential of emerging technologies in a controlled environment. The name "Prometheus'' meaning "forethought" is fitting because this project is all about planning for the future, applying and testing socio-technical standards and exploring the potential alignment of new technologies before they are widely adopted.

The U.K. was the first jurisdiction to launch a regulatory sandbox in 2015, and since then, over twenty jurisdictions have actively implemented or explored the concept. According to the White House AI R&D Strategic Plan 2023 update "Appropriate programs should be established for academic and industrial researchers to conduct research within secured and curated testbed environments established by federal agencies ... Industry and academia are the primary sources for emerging AI systems. Promoting and coordinating R&D subject matter expert participation in standards and benchmarking activities are critical." Similarly, the first drafts of the proposed EU AI Act also call for "sandbox" creation (e.g., in Title V).

By creating a regulatory sandbox for AI and AIS, the project can provide a space for regulators, manufacturers, innovators, and the public to work together collaboratively and experimentally to test out the viability of the various stages of deployment before it is released upon the world.

A key benefit of a regulatory sandbox is that stakeholders can test innovative products, technologies, laws and business models related to AIS using the same interoperable socio-technical standards. This live, time-bound testing can provide valuable insights into how emerging technologies can comply with existing regulations and how new regulatory frameworks might need to be adjusted to support them. In this way, the sandbox can provide an opportunity for incumbents and challengers alike to experiment with new ideas outside of existing regulatory frameworks, which can lower the barriers to entry and reduce the cost and risk of innovation, testing the compliance of AIS with potential laws, rules and regulations while also testing for the fitness of the proposed laws. This process removes ambiguity, streamlines the path to adoption, and enables regulators to be proactive rather than reactive in their approach to regulating emerging technologies.

The Prometheus Project aspires to kickstart a smarter global regulatory framework for increasingly autonomous AI systems.

For more information please contact:

The Spatial Web Foundation

inquiries@thespatialweb.org

VERSES

inquiries@verses.ai

Dentons

peter.stockburger@dentons.com

Authors

Gabriel René, Philippe Sayegh, Capm Petersen, Chase Pletts[1]

Dan Richardson[2]

Peter Stockburger, Eric Tanenblatt[3]

Acknowledgements:

Dan Mapes, Don Moody, Karl Friston, Maxwell Ramstead, George Percivall, Mahault Albarracin, Safae Essafi Tremblay, Eric Holder, Ines Hipolito, Steven Swanson, Michael Wadden, Bastiaan den Braber

About the Spatial Web Foundation

The Spatial Web Foundation (SWF) is a non-profit organization dedicated to the development and ethical use of technology, specifically in the creation and implementation of the Spatial Web protocol. The foundation's core initiatives include developing open standards and protocols, promoting interoperability, and educating policymakers and the public. SWF supports transparency and accountability in the development and use of technology, promotes responsible innovation, engages in dialogue and collaboration with various stakeholders, and adopts a proactive approach to addressing emerging ethical challenges. Ultimately, SWF seeks to create a more inclusive, accessible, and equitable internet that empowers individuals and communities to connect, create, and thrive. For more information visit: spatialwebfoundation.org

About VERSES

VERSES is a cognitive computing company specializing in next-generation Artificial Intelligence. Modeled after natural systems and the design principles of the human brain and the human experience. Built on open Spatial Web standards (IEEE P2874) our platform transforms disparate data into knowledge models that foster trustworthy collaboration between humans, machines, and AI, across digital and physical domains. Imagine a smarter world that elevates human potential through innovations inspired by nature. Learn more at VERSES, LinkedIn, and Twitter.

About Dentons

Dentons is designed to be different. As the world's largest global law firm with 21,000 professionals in over 200 locations in more than 80 countries, Dentons can help you grow, protect, operate and finance your business. Dentons polycentric and purpose-driven approach, together with their commitment to inclusion, diversity, equity and ESG, ensures they challenge the status quo to stay focused on what matters most to you. For more information visit Dentons.com

---

1 **VERSES**
Gabriel Rene, Founder and CEO; Philippe Sayegh, Chief Adoption Officer; Capm Petersen, Chief Innovation Officer; Chase Pletts, Director of Market Intelligence

2 **The Spatial Web Foundation**
Dan Richardson, Director of Market Analysis

3 **Dentons**
Peter Stockburger, managing partner of Dentons' San Diego office, co-lead of the firm's Global Autonomous Vehicles Practice.
Eric Tanenblatt, Dentons' Global Chair of Public Policy and Regulation. He also leads the firm's US Public Policy Practice, and co-leads the firm's Global Autonomous Vehicles Practice.