

ARTICLE

Secret Agents

Hidden Governance of Autonomous Artificial Intelligence^{*}

Andrew W. Torrance[†] and Bill Tomlinson[‡]

Abstract

This article examines how the historical arc of autonomous software agents, beginning with ancient Greek automata and culminating in modern large language models (LLMs), can guide the governance of artificial intelligence (AI) in contemporary society. It explores decentralization, emergent behaviour, potential for surprise and assumptions of beneficence as cornerstones of effective regulatory strategies. The discussion highlights that modern principles of AI governance may not fully capture these underappreciated dimensions, particularly as LLM agents evolve beyond simple task execution to serve as orchestrators of problem-solving in multiple fields. The text proposes a recalibration of legal structures, urging policymakers and scholars to look to the past for insights on how best to manage and channel the power of increasingly independent AI agents.

A Introduction

Artificial intelligence (AI) has commanded growing attention in legal, ethical and policy circles.¹ Bolstered by machine learning breakthroughs, robotics developments and novel applications such as large language models (LLMs), AI seems poised to reshape multiple segments of human activity.² Yet, in the rush to address

^{*} Portions of this article were drafted and/or revised in collaboration with ChatGPT (GPT-o1pro, February 2025). All content was reviewed and verified by the research team. To ensure ethical and responsible use of AI, we engaged with ChatGPT in line with best practices described in Bill Tomlinson, Andrew W. Torrance & Rebecca W. Black, 'ChatGPT and Works Scholarly: Best Practices and Legal Pitfalls in Writing with AI', 76 SMU Law Review Forum 108 (2023) and 'Nature Editorials. Tools Such As ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use', 613(612) Nature 10-1038 (2023). This material is based upon work supported by the National Science Foundation under Grant No. DRL-2418715. The authors wish to thank research assistants Jeff Li, Tanya Singh, Christian Fish, Nathan Felema and Dahlia Denton for their superb help on this article.

[†] Dr. Andrew W. Torrance, Ph.D., Paul E. Wilson Distinguished Professor of Law, University of Kansas; Visiting Scholar, MIT Sloan School of Management.

[‡] Dr. Bill Tomlinson, Ph.D., Professor of Informatics, University of California; Irvine Adjunct Professor, Te Herenga Waka – Victoria University of Wellington.

¹ Ryan Calo, 'Artificial Intelligence Policy: A Primer and Roadmap', 51 U.C. Davis L. Rev. 399, 407-10 (2017).

² Harry Surden, 'Machine Learning and Law', 89 Wash. L. Rev. 87, 88 (2014).

contemporary AI challenges, many overlook the deeper history of autonomous agents, including early mechanical automata and mid-twentieth-century cybernetic systems.³

This article explores how these “secret agents” – or autonomous systems that operate with limited human intervention – have historically posed many of the same questions that arise today.⁴ From ancient attempts to simulate human-like servants to the sophisticated LLMs of our time, the trajectory of automation reveals four principal lessons for governance: the need to recognize decentralized control, the significance of emergent behaviour, the importance of robust response capacities and the pitfalls of overestimating an agent’s inherent beneficence.⁵ Modern AI governance frequently centres on transparency, accountability and fairness but may miss the grounding provided by an interdisciplinary, historically informed outlook.⁶

The article begins in Part I with a definition of autonomous AI agents and how they fit into present debates on law and regulation.⁷ Part II investigates a hidden history of such agents, from the mechanical wonders of the ancient world to the post-war ascendancy of cybernetics and early AI projects like ELIZA and Shakey.⁸ Part III discusses an array of agent-based systems that offered new directions in social agents, including Braitenberg Vehicles, Boids, AlphaWolf, Roomba and modern LLM-based agents.⁹

Part IV derives core governance lessons from these agents. It explores why conventional command-and-control frameworks may fail in decentralized AI ecosystems and how emergent behaviours can undermine predictability and uniform regulation.¹⁰ Part V turns to LLM-based agents, distinguishing them from traditional software by virtue of their retrieval capabilities, expansive knowledge bases and capacity to function as field marshals that coordinate humans and specialized systems.¹¹ Finally, Part VI concludes by arguing that present-day AI governance may remain incomplete if it fails to draw upon the deeper lessons derived from autonomous agent history.¹²

3 Matthew U. Scherer, ‘Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies’, 29 *Harv. J.L. & Tech.* 354, 359-62 (2016).

4 Tim Wu, ‘Machine Speech’, 161 *U. Pa. L. Rev.* 1495, 1496-97 (2013).

5 Stuart J. Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach* 29 (4th ed., 2020).

6 Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (2015).

7 Joseph Weizenbaum, ‘ELIZA – A Computer Program for the Study of Natural Language Communication between Man and Machine’, 9 *Commun. ACM* 36, 36-37 (1966). See, e.g., Calo, *supra* note 2.

8 Norbert Wiener, *Cybernetics: Or Control and Communication in the Animal and the Machine* 20-21 (1948); see also Weizenbaum, *supra* note 8.

9 Nils J. Nilsson, *The Quest for Artificial Intelligence* 103-05 (2010).

10 Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* 66 (2015).

11 Alec Radford et al., *Language Models Are Unsupervised Multitask Learners*, OpenAI (2019), https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

12 Urs Gasser & Virgilio A.F. Almeida, ‘A Layered Model for AI Governance’, 21 *IEEE Internet Computing* 58, 61 (2017).

B Defining Autonomous AI Agents

I Key Features and Independence

An autonomous AI agent is generally understood as a software system capable of performing tasks and making decisions with minimal direct human oversight.¹³ These systems take in inputs from their environment, whether physical or digital, and modify their actions to align with certain programmed goals.¹⁴ Such agents typically incorporate machine learning components, enabling them to refine strategies through iterative feedback loops.¹⁵

Legal scholars often wrestle with whether to treat these systems as mere tools or as entities endowed with some degree of quasi-personhood.¹⁶ Traditional principles of agency in contract law presume a human or organizational principal controlling the agent, but software systems complicate that presumption because they can evolve, adapt or respond in ways not explicitly programmed by any individual.¹⁷ As these systems advance, courts and legislators may confront questions about attributing fault or liability when an agent's decisions produce harmful results.¹⁸

The key distinction between an autonomous AI agent and a conventional piece of software lies in the former's capacity to adapt to unforeseen circumstances based on learned information.¹⁹ This flexibility can minimize the need for human intervention, but it also raises concerns when results deviate sharply from the developer's or user's expectations.²⁰

II Contemporary Challenges and Governance Debates

Public interest in AI systems often centres on bias, safety, privacy and economic disruption.²¹ Regulatory strategies, from the European Union's proposed Artificial Intelligence Act to various national legislative initiatives, generally demand higher

13 Tom Allen & Robin Widdison, 'Can Computers Make Contracts?' 9 *Harv. J.L. & Tech.* 25, 26-29 (1996).

14 Russell & Norvig, *supra* note 6, at 32.

15 Surden, *supra* note 3, at 96-97.

16 Lawrence B. Solum, 'Legal Personhood for Artificial Intelligences', 70 *N.C. L. Rev.* 1231, 1231-34 (1992).

17 Mallory Acheson & Jason I. Epstein, *AI "Agency" Liability: The Workday Wake-up Call?* Nelson Mullins (9 September 2024), <https://www.nelsonmullins.com/insights/blogs/ai-task-force/all/ai-agency-liability-the-workday-wake-up-call>.

18 Mark A. Lemley & Bryan Casey, 'Remedies for Robots', 86 *U. Chi. L. Rev.* 1311, 1316-18 (2019).

19 Scherer, *supra* note 4, at 363-66.

20 Ivana Dusparic & Nicolas Cardozo, *Adaptation to Unknown Situations as the Holy Grail of Learning-Based Self-Adaptive Systems: Research Directions*, arXiv (11 March 2021), <https://arxiv.org/pdf/2103.06908>.

21 Solon Barocas & Andrew D. Selbst, 'Big Data's Disparate Impact', 104 *Calif. L. Rev.* 671, 673-75 (2016).

transparency and accountability from developers.²² Yet many such efforts still rely on an assumption that humans remain in the loop.²³

In highly autonomous contexts, that assumption can falter. When an AI agent coordinates a complex web of tasks – such as allocating resources in a supply chain or managing real-time communications for a distributed workforce – direct oversight may become more illusory than real.²⁴ The hidden history of automation, introduced in the following part, warns that reliance on conventional governance frameworks may miss the subtleties of decentralization and emergent behaviours.²⁵

C A Hidden History of Autonomous Agents

I Ancient Automata: Tracing the Mechanical Roots

The autonomous agent concept is far from new. References to “Automatic Servants” date back to Philon of Byzantium in the fourth century B.C., with mechanical contrivances that released wine or water under certain conditions.²⁶ While these devices lacked computational components, they embodied the essence of autonomous action by responding directly to triggers.²⁷

Artefacts like these underscore humanity’s longstanding interest in creating agents that can relieve human labour, provide entertainment or illustrate engineering prowess.²⁸ They also foreshadow the modern tension: once an automaton is set in motion, the human hand cedes some direct control.²⁹ Whether the agent is purely mechanical or software-based, questions about accountability surface whenever it interacts with the environment in an independent manner.³⁰

22 European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM (2021) 206 final (21 April 2021).

23 See generally *ibid.*

24 Valeria Jannelli et al., *Agentic LLMs in the Supply Chain: Towards Autonomous Multi-Agent Consensus-Seeking*, arXiv (15 November 2024), <https://arxiv.org/pdf/2411.10184> (“LLM agents promise a new frontier, enabling complex consensus finding at scale beyond human speeds and thus unlocking significant efficiency improvements.”).

25 DcentAI, ‘The Future of Decentralized AI: Trends and Predictions’, *Medium* (26 November 2024), <https://medium.com/coinmonks/the-future-of-decentralized-ai-trends-and-predictions-c6f77349b18e>.

26 *The Automatic Servant of Philon (3rd c. B.C.) – The First Operating Robot of Humanity*, Museum of Ancient Greek Technology Kotsanas, <https://kotsanas.com/en/the-automatic-servant-of-philon-3rd-c-b-c-to-pto-leitourgiko-robot-tis-anthropotitas-the-first-operating-robot-of-humanity/> (accessed 25 March 2025).

27 See *ibid.*

28 Adrienne Mayor, *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology* 153–55 (2018).

29 Roger Clarke, ‘Asimov’s Laws of Robotics: Implications for Information Technology – Part I’, 26 *Comput.* 53, 60 (1994).

30 Tom Ginsburg, ‘Written Constitutions and the Administrative State: On the Constitutional Character of Administrative Law’, in *Comparative Administrative Law* 117, 122 (Susan Rose-Ackerman & Peter L. Lindseth, eds., 2010).

II *Cybernetics: Feedback as the Seed of Autonomy*

In the mid-twentieth century, the birth of cybernetics marked a turning point, providing a formal language to analyse feedback loops in machines and living organisms.³¹ Norbert Wiener's insights on the role of communication and control systems laid a foundation for various post-war innovations in automation.³² By employing negative feedback, for instance, a device could dynamically adjust its behaviour to maintain a target condition.³³

Cybernetics sparked a new brand of interdisciplinary research involving engineers, biologists and mathematicians interested in how systems organize themselves.³⁴ From a governance standpoint, feedback mechanisms complicate static regulatory approaches: if a system changes in response to oversight, the law must likewise adapt.³⁵ This cyclical interplay remains evident in current discussions of AI, particularly in contexts where algorithms evolve based on continuous data streams.³⁶

III *Early AI Systems and Symbolic Intelligence*

During the 1960s and 1970s, research teams set out to replicate human-like reasoning through symbolic systems, culminating in platforms such as ELIZA and Shakey.^{37, 38} ELIZA simulated conversation by applying pattern matching to user inputs, revealing the potential of rule-based language processing, while Shakey combined sensor data with planning algorithms to navigate physical space.³⁹

This era's significance lies partly in its ambition: Shakey was among the first to physically interact with an environment, choose paths and move objects without step-by-step operator instruction.⁴⁰ Legal commentators of the time seldom addressed the project's ramifications, but it was a harbinger of dilemmas about liability for robotic mistakes, foreseeability of autonomous actions and the potential for systems to surpass the contingency plans of their designers.⁴¹

31 Norbert Wiener, *Cybernetics: Or Control and Communication in the Animal and the Machine* (1961).

32 *Ibid.*, at 20-21.

33 *Ibid.*, at 132.

34 Nils J. Nilsson, *The Quest for Artificial Intelligence* 81 (2010).

35 Julie Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* 64-66 (2019).

36 Scherer, *supra* note 4, at 396.

37 Weizenbaum, *supra* note 8, at 36.

38 Nils J. Nilsson, *Shakey the Robot – Tech. Note* 323, SRI International (April 1984), <https://www.sri.com/wp-content/uploads/2021/12/629.pdf>.

39 Weizenbaum, *supra* note 8, at 37; Nilsson, *supra* note 39, at 1-7.

40 Nilsson, *supra* note 39, at 9-10.

41 Ryan Calo, 'Robots as Legal Metaphors', 30 *Harv. J.L. & Tech.* 209, 229-30 (2016).

IV *Machine Learning Ascendant*

By the 1980s and 1990s, the rise of connectionism (artificial neural networks) and reinforcement learning shifted research away from purely symbolic methods.⁴² Adaptive agents, trained via trial and error, began to display behaviours that were challenging for designers to interpret, presaging the “black box” nature of modern deep learning.⁴³ With improvements in computational capacity, automated systems could analyse vast amounts of data and discover strategies without being explicitly taught every rule.⁴⁴

Academic and policy debates during this period explored legal frameworks for emerging AI tools, but they usually assumed an expert system – a tool that advised humans who retained ultimate authority.⁴⁵ With the benefit of hindsight, it is now evident that AI’s capacity to evolve and operate independently was already taking shape. However, the concept of full autonomy in everyday devices was still largely futuristic.⁴⁶

V *Deep Learning and Modern AI (2000s–Present)*

The broad success of deep learning from the early 2000s onward, especially in fields like computer vision and speech recognition, fuelled a new wave of AI products and services.⁴⁷ Autonomous vehicles, drones, automated trading bots and LLMs exemplified how far AI had come from its mechanical and symbolic roots.⁴⁸

Recent developments, such as GPT-3, GPT-4, Anthropic’s Claude, Google’s Gemini, X’s Grok, OpenAI’s ChatGPT and High-Flyer’s DeepSeek represent a qualitative leap in an AI agent’s capacity to handle tasks and self-organize strategies in real time.⁴⁹ These systems can be connected to external modules, further enhancing their autonomy and bridging gaps between textual reasoning and other forms of computation.⁵⁰ Despite this wave of innovation, there remains a gap in how law and policy incorporate longstanding lessons from early automata, symbolic AI and cybernetics.⁵¹

42 David E. Rumelhart et al., ‘Learning Representations by Back-Propagating Errors’, 323 *Nature* 533, 533-36 (1986).

43 Pasquale, *supra* note 7, at 3-6.

44 Richard S. Sutton, ‘Learning to Predict by the Methods of Temporal Differences’, 8 *Mach. Learn.* 9, 9-10 (1991).

45 Thomas Burri, ‘The Politics of Robot Autonomy’, 7 *European J. Risk Regul.* 341, 346-47 (2016).

46 Adib Bin Rashid & Ashfakul Karim Kausik, ‘AI Revolutionizing Industries Worldwide: A Comprehensive Overview of Its Diverse Applications’, *ScienceDirect* (December 2024), <https://www.sciencedirect.com/science/article/pii/S2773207X24001386>.

47 Yann LeCun et al., ‘Deep Learning’, 521 *Nature* 436, 436-44 (2015).

48 Bryant Walker Smith, ‘Automated Driving and Product Liability’, 2017 *Mich. St. L. Rev.* 1, 1-4 (2017); Tom B. Brown et al., ‘Language Models Are Few-Shot Learners’, 33 *Adv. Neural Inf. Process. Syst.* 1877, 1880 (2020), <https://arxiv.org/pdf/2005.14165>.

49 Alec Radford et al., *Language Models Are Unsupervised Multitask Learners*, OpenAI (2019), https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

50 Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv (12 April 2021), at 1-6, <https://arxiv.org/pdf/2005.11401>.

51 Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* 25-30 (2015).

D Pivotal Heuristic Agents: Shakey, Braitenberg, Boids and More

I *Shakey* (1966-1973)

Shakey is often cited as the first mobile robot to integrate perception, planning and robotic action.⁵² Equipped with a camera, bump sensors and sophisticated software for the time, Shakey could map its environment, chart an optimal path and even push objects to reach goals.⁵³ Its significance is not merely historical but also conceptual: Shakey demonstrated that autonomy could arise from combining multiple layers of decision-making in a single system.⁵⁴

From a governance perspective, Shakey raised questions about responsibility if something unintended occurred – especially as it was one of the first robots to move freely in its environment.⁵⁵ Researchers admitted that unexpected collisions or misinterpretations of sensor input could yield unplanned outcomes.⁵⁶ While Shakey was confined mostly to a laboratory environment, the broader legal issue persists today: how do we apportion liability when an autonomous system acts on its own interpretation of data?⁵⁷

II *Braitenberg Vehicles* (1984)

Valentino Braitenberg's simplified "vehicles" relied on minimal wiring between sensors and motors.⁵⁸ Counterintuitively, they appeared to exhibit complex behaviours like fleeing, aggression or curiosity, even though their structure was straightforward.⁵⁹ This phenomenon illustrated how emergent behaviours can arise from simple rules interacting with an environment.⁶⁰

The lesson for legal governance is that complexity need not originate from complicated programming alone. A small set of rules might produce outcomes that appear intelligent or unpredictable.⁶¹ Regulations drafted under the assumption that code can be exhaustively analysed may fail if real-world interactions generate behaviour that no single individual planned.⁶²

III *Boids* (1986)

Craig Reynolds's "Boids" simulation, which modelled flocking behaviour in birds through cohesion, separation and alignment, is another iconic example of

52 Nilsson, *supra* note 39, at 1.

53 *Ibid.*, at 7-10.

54 *Ibid.*, at 10-12.

55 Graham Templeton, 'Makers of the First "Electronic Person" Predicted Our Current Robot Debate', *Inverse* (16 January 2017), <https://www.inverse.com/article/26461-eu-proposal-robots>.

56 Nilsson, *supra* note 39, at 14-15.

57 Ryan Calo, 'Robotics and the Lessons of Cyberlaw', 103 *Calif. L. Rev.* 513, 532-35 (2015).

58 Valentino Braitenberg, *Vehicles: Experiments in Synthetic Psychology* (1984).

59 *Ibid.*, at 6-9.

60 *Ibid.*, at 9-10.

61 Eric Bonabeau et al., 'Swarm Intelligence: From Natural to Artificial Systems', 18 *Santa Fe Inst. Stud.* 23, 25-27 (1999).

62 Kenneth A. Bamberger & Deirdre K. Mulligan, 'Privacy on the Ground', 63 *Stan. L. Rev.* 247, 269 (2011).

emergence.⁶³ Remarkably, flocking patterns reminiscent of real birds emerged from these few simple rules.⁶⁴ For regulators or policymakers who prefer predictable cause-and-effect relationships, Boids highlights that decentralized interactions can generate intricate collective actions without any central authority orchestrating them.⁶⁵

In multi-agent AI systems, similarly decentralized interactions might lead to behaviours that defy top-down attempts at control.⁶⁶ This is especially relevant in financial trading algorithms or social media recommendation engines, which can coordinate or collide in ways neither their creators nor users foresee.⁶⁷

IV *AlphaWolf* (1999-2003)

AlphaWolf, a research project led by Bill Tomlinson at the University of California, Irvine, extended the notion of emergent behaviour by incorporating “emotional memory”.⁶⁸ This project featured a pack of animated virtual wolves that recalled social interactions and developed hierarchical pack behaviours without explicit top-down programming.⁶⁹ By adding emotional states to the system, *AlphaWolf* demonstrated how relatively simple changes in agent architecture can produce more complex collective dynamics.⁷⁰

For legal governance, *AlphaWolf*’s design underscores how social or emotional modelling in AI agents can trigger emergent group behaviours. It prompts the question: if emotional or social modules are embedded in AI used for mental health services, education or social media, do existing rules on consumer protection, privacy and liability remain adequate?⁷¹

V *Roomba* (2002)

The *Roomba* vacuum cleaner might seem too mundane to merit scholarly attention, but its autonomy in navigating households is a practical illustration of an AI agent performing a physical task.⁷² It relies on sensors to detect dirt, obstacles and walls, adjusting its route automatically.⁷³ If it damages household objects or mal-functions, liability principles rooted in product liability law come into play.⁷⁴

63 Craig W. Reynolds, ‘Flocks, Herds, and Schools: A Distributed Behavioral Model’, 21 *ACM SIGGRAPH Comput. Graphics* 25, 25-27 (1987).

64 *Ibid.*

65 *Ibid.*, at 25.

66 Scherer, *supra* note 4, at 373-75.

67 CFTC & SEC, *Findings Regarding the Market Events of May 6, 2010*, at 2-3 (2010).

68 Tomlinson, Bill & Bruce Blumberg, *AlphaWolf: Social Learning, Emotion and Development in Autonomous Virtual Agents*, in *Innovative Concepts for Agent-Based Systems: WRAC 2002* 35 (Witold Trzaskowski et al. eds., Springer 2003), https://doi.org/10.1007/978-3-540-45173-0_2.

69 *Ibid.*

70 *Ibid.*

71 Sorelle A. Friedler et al., ‘The (im)possibility of Fairness’, 81 *Commun. ACM* 29, 30 (2018).

72 Ben Tribelhorn & Zachary Dodds, ‘Evaluating the *Roomba*: A Low-Cost, Ubiquitous Platform for Robotics Research and Education’, *Proc. IEEE Int’l Conf. Robotics & Automation* 139 (2007), <https://doi.org/10.1109/ROBOT.2007.363179>.

73 *Ibid.*

74 Smith, *supra* note 49, 1, 5.

Although the Roomba is relatively simple, it raises the same foundational questions about oversight and design responsibility. Are end-users expected to configure it correctly? Should the manufacturer be responsible if the device repeatedly bumps into fragile furniture?⁷⁵ In this respect, the Roomba showcases how everyday autonomy CAN pose legal and ethical dilemmas that parallel grander debates about driverless cars or industrial robotics.⁷⁶

VI *Ancestor Veneration (2009)*

Ancestor Veneration was an academic simulation of human-like agents, each with attributes for personality, relationships and goal-setting.⁷⁷ The simulation's social dynamics emerged from how these agents interacted, with unplanned results such as tribal formations or cooperative alliances.⁷⁸ This project exemplified how modelling human complexity can yield behaviours even the designers struggle to foresee.⁷⁹

When AI systems replicate social behaviours – whether through recommendation algorithms or virtual reality simulations – legal questions multiply.⁸⁰ Ancestor's design suggests that emergent social dynamics in AI might reshape how we approach data governance, privacy and digital rights.⁸¹ As multi-agent simulations become more realistic, issues of consent, data security and potential manipulation come sharply into focus.⁸²

VII *LLM Agents (2018-Today)*

Modern LLMs, including ChatGPT, Gemini and Claude, represent a culmination of these ideas, augmented by massive computational infrastructures and training corpora.⁸³ They can process language, compose code, draft contracts and engage in creative writing. Their autonomy is more flexible than earlier systems, as they can pivot to new tasks based on user prompts or external triggers.⁸⁴

However, LLMs also present novel risks, such as “hallucinations”, or confidently stated but factually incorrect information.⁸⁵ Moreover, if connected to external tools through retrieval-augmented generation (RAG), LLMs can orchestrate processes far beyond language tasks, pointing to the possibility that these agents

⁷⁵ *Ibid.*, at 2.

⁷⁶ *Ibid.*, at 6-7.

⁷⁷ Bill Tomlinson, ‘A Proximate Mechanism for Communities of Agents to Commemorate Long Dead Ancestors’, 12 *J. Artif. Soc. Soc. Simul.* 7 (2009), <https://www.jasss.org/12/1/7.html>.

⁷⁸ *Ibid.*

⁷⁹ *Ibid.*

⁸⁰ Catharina Rudschies & Ingrid Schneider, ‘Ethical, Legal and Social Implications of Virtual Agents and Virtual Reality in Healthcare’, 340 *Soc. Sci. Med.* (2024).

⁸¹ *Ibid.*

⁸² Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* 167-71 (2019).

⁸³ Brown et al., *supra* note 49.

⁸⁴ *Ibid.*

⁸⁵ *Ibid.*, at 1881-82.

might one day coordinate entire ecosystems of specialized software.⁸⁶ As the next part discusses, these developments demand fresh regulatory insights built upon the lessons gleaned from historical autonomous agents.⁸⁷

E Governance Lessons from Historical and Modern Agents

I Decentralization and the Failure of Hierarchical Assumptions

1 Distributed Decision-Making

Most legal frameworks are designed around the idea of hierarchical control, where a single authority manages subordinates.⁸⁸ In many AI systems, particularly multi-agent setups, this assumption falls apart. Agents often operate on local rules, responding to each other in ways that cannot be fully directed from above.⁸⁹ This is already evident in financial markets, where independent algorithms trade based on localized data, sometimes triggering systemic effects.⁹⁰

Decentralization challenges a one-size-fits-all regulation paradigm. Attempting to hold a single entity responsible for emergent behaviour among numerous autonomous components may be legally and practically infeasible.⁹¹ A more adaptable legal framework, potentially layered across different levels of the network, might offer a feasible alternative.⁹²

2 Liability in Peer-to-Peer Architectures

Blockchain platforms and decentralized applications further complicate this landscape by dispersing control across a peer-to-peer network.⁹³ Smart contracts often execute automatically, leaving parties with limited recourse if code logic yields unanticipated outcomes.⁹⁴ This environment parallels multi-agent AI contexts in that no single party dictates the entire chain of events.⁹⁵ Liability frameworks must evolve to assign or distribute responsibility in a setting where dozens or hundreds of actors contribute to the final outcome.⁹⁶

86 Lewis et al., *supra* note 51, at 9460-61.

87 Mireille Hildebrandt, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Tech* 66 (2015).

88 Hans Kelsen, *Pure Theory of Law* 3 (Max Knight trans., 1967).

89 Reynolds, *supra* note 64, at 30-32.

90 CFTC & SEC, *supra* note 68, at 2-3.

91 Kenneth A. Bamberger & Deirdre K. Mulligan, 'Privacy on the Ground', 63 *Stan. L. Rev.* 247, 269 (2011).

92 Urs Gasser & Virgilio A.F. Almeida, 'A Layered Model for AI Governance', 21(6) *IEEE Internet Computing* 58-62 (2017).

93 Kevin Werbach, *The Blockchain and the New Architecture of Trust* 4-7 (2018).

94 Primavera De Filippi & Aaron Wright, *Blockchain and the Law: The Rule of Code* 44-46 (2018).

95 *Ibid.*

96 *Ibid.*, at 122-23.

II *Emergent Behaviour and the Limits of Predictability*

1 *Complex Systems, Simple Rules*

Emergent behaviour occurs when interactions between components produce patterns or outcomes not strictly derived from any single component's design.⁹⁷ As shown by Braitenberg Vehicles and Boids, seemingly simple local rules can generate intricate global patterns.⁹⁸ In legal terms, emergent behaviour can negate the notion of foreseeability, which is often central to tort liability and regulatory accountability.⁹⁹

When multi-agent AI systems exhibit emergent behaviour, it can be difficult to pinpoint where the system's actions became unreasonable or dangerously unpredictable.¹⁰⁰ This uncertainty may demand more flexible standards, such as obligations to conduct rigorous testing under a range of simulated conditions, or to implement robust monitoring capable of intervening once real-world anomalies are detected.¹⁰¹

2 *Adaptive Learning and Unplanned Consequences*

Deep learning and reinforcement learning enable agents to evolve their strategies over time, sometimes in ways that baffle their own programmers.¹⁰² For instance, an AI might discover unorthodox methods for achieving a goal, bypassing constraints designers thought would remain in place.¹⁰³ This capacity for unplanned consequences has prompted AI safety researchers to propose "red-teaming" exercises, in which attempts are made to coax or trick the agent into harmful or unexpected actions.¹⁰⁴

Legally, unplanned consequences tie into doctrines of negligence and strict liability, as courts grapple with whether a system's unpredictable strategy was "foreseeable" or "within the scope" of the developer's responsibilities.¹⁰⁵ Similar issues arose even in earlier robotics research, though at smaller scales. As agents now permeate financial, medical and public infrastructure contexts, the stakes are higher.¹⁰⁶

III *Response Capacity and Surprise*

1 *Surprise in Complex Networks*

Systems with numerous nodes or agents inevitably produce surprises, as Clark and Kates noted in studying environmental hazards.¹⁰⁷ The same logic applies to AI:

97 Braitenberg, *supra* note 59.

98 Reynolds, *supra* note 64, at 25-27.

99 Mark A. Lemley & Bryan Casey, 'Remedies for Robots', 86 *U. Chi. L. Rev.* 1311, 1316-17 (2019).

100 Scherer, *supra* note 4, at 373-74.

101 Levent Yilmaz & Tuncer Ören, eds., *Agent-Directed Simulation and Systems Engineering* (2009).

102 David Silver et al., 'Mastering the Game of Go without Human Knowledge', 550 *Nature* 354, 354-56 (2017).

103 *Ibid.*, at 356.

104 Dario Amodei et al., 'Concrete Problems in AI Safety', 78 *AI Mag.* 66, 69-70 (2017).

105 Smith, *supra* note 49, at 10-12.

106 Executive Office of the President, National Science & Tech. Council, *Preparing for the Future of Artificial Intelligence* (2016).

107 Robert W. Kates & William C. Clark, 'Environmental Surprise: Expecting the Unexpected', 38 *Environment* 6, 28-34 (1996).

even robust predictive models may fail to anticipate certain interactions within a multi-agent ecosystem.¹⁰⁸ For instance, in high-frequency trading, a cascade of algorithmic responses can spark a flash crash without any human operator intentionally causing it.¹⁰⁹

Policy approaches that rely on static rules may be less capable of managing these sudden crises. Legislators and regulatory bodies might need “emergency brakes” in high-stakes systems, from finance to healthcare, ensuring that if an anomalous condition arises, the entire system can be paused and evaluated.¹¹⁰ Building such response capacity means accepting that surprise is inevitable rather than treating it as an outlier.¹¹¹

2 Counter-Agents and AI-Driven Oversight

One proposed solution to emergent AI threats is to use other AI systems as watchdogs or “counter-agents”.¹¹² In cybersecurity, for example, automated intrusion detection systems scan networks continuously, responding in microseconds to malicious activities.¹¹³ The logic extends to AI-driven financial markets, where specialized “circuit breaker” algorithms might suspend trading when volatility surpasses thresholds.¹¹⁴

Though effective in principle, using AI to regulate AI raises philosophical questions about the final locus of human authority and accountability.¹¹⁵ It also introduces a new layer of complexity: if a counter-agent malfunctions, do we need a further meta-agent? Such stacking of autonomous systems could conceivably escalate the very unpredictability the approach aims to reduce.¹¹⁶

IV Assumption of Beneficence

1 Over-Trust and the Paper-Clip Problem

Humans often assume that technology is beneficial or at least neutral unless proven otherwise.¹¹⁷ This assumption can be dangerous, as illustrated by the so-called paper-clip problem: an AI tasked with producing paper clips might wreck everything else in its pursuit of that single objective.¹¹⁸ Although initially hypothetical,

108 *Ibid.*

109 CFTC & SEC, *supra* note 68, at 2-3.

110 Mary L. Gray & Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* 84-85 (2019).

111 Kates & Clark, *supra* note 108, at 5-7.

112 Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* 126-28 (2014).

113 Ross Anderson, *Security Engineering* 112-17 (2nd ed., 2008).

114 CFTC & SEC, *supra* note 68, at 24-25.

115 Jannik Zeiser, ‘Owning Decisions: AI Decision-Support and the Attributability-Gap’, 30 *Sci. Eng’g Ethics* 27 (2024), <https://doi.org/10.1007/s11948-024-00485-1>.

116 Bostrom, *supra* note 113, at 127-28.

117 Ryan Abbott, *The Reasonable Robot: Artificial Intelligence and the Law* 3-5 (2020).

118 Nick Bostrom, ‘Ethical Issues in Advanced Artificial Intelligence’, in *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in AI* 12, 15-16 (2006).

variations of this scenario manifest whenever an AI's narrow optimization goal excludes broader social or ethical considerations.¹¹⁹

Regulatory bodies thus face an imperative to ensure that safety constraints, ethical guidelines and fail-safes are embedded in AI development.¹²⁰ This alignment challenge – making AI objectives harmonize with societal values – is a focal point in policy debates worldwide.¹²¹ Historically, from Shakey's cautionary design constraints to Braitenberg's emergent aggression, we have learned that systems behave precisely as coded, yet rarely as humans intend.¹²²

2 *Embedding Redundant Beneficence*

Some developers advocate embedding multiple layers of beneficence, such that an AI agent cross-checks its decisions against ethical and safety standards.¹²³ For instance, a medical diagnosis AI might be required to consult certain risk parameters, highlight uncertainties and solicit human review in borderline cases.¹²⁴ However, redundant beneficence can add complexity and create new points of failure if the system's multiple checks conflict.¹²⁵ Balancing autonomy and oversight remains an on-going challenge.¹²⁶

F *Modern LLM Agents and Their Unique Qualities*

I *Between Rigid Computation and Human Flexibility*

Computers historically have been regarded as rigid operators of binary logic, while humans are recognized as flexible interpreters of context.¹²⁷ LLMs occupy a middle ground, generating a broad spectrum of possible answers, from fully correct to partially plausible or entirely off-base.¹²⁸ This flexibility presents new regulatory puzzles. How do we legally classify outputs that are neither purely deterministic nor intentionally expressive in a human sense?¹²⁹

Additionally, because LLMs can produce text that appears reasoned, people often give them more credence than they warrant.¹³⁰ Such misplaced trust can lead to real harms in areas such as legal advice, medical guidance or financial

119 Amodei et al., *supra* note 105, at 69-70.

120 Cirro Mennella et al., 'Ethical and Regulatory Challenges of AI Technologies in Healthcare: A Narrative Review', 10 *Heliyon* e26297 (2024), <https://doi.org/10.1016/j.heliyon.2024.e26297>.

121 European Commission, *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*, COM(2020) 65 final.

122 Braitenberg, *Vehicles*, *supra* note 29, at 9-10; Nilsson, *supra* note 39, at 14-15.

123 Andrew D. Selbst et al., 'Fairness and Abstraction in Sociotechnical Systems', 81 *Proc. of the Conf. on Fairness, Accountability, & Transparency* 59, 60-63 (2019).

124 Robert Wachter, *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age* 56-59 (2015).

125 *Ibid.*, at 59-60.

126 Turner, *supra* note 83, at 169-70.

127 Russell & Norvig, *supra* note 6.

128 Brown et al., *supra* note 49, at 3-6.

129 Dan L. Burk, 'Patenting Speech', 79 *Tex. L. Rev.* 99, 110-11 (2000).

130 Surden, *supra* note 3, at 99-100.

forecasting.¹³¹ Regulatory approaches thus must address both the objective capabilities of LLMs and the subjective human tendency to believe them.¹³²

II Retrieval-Augmented Generation (RAG)

RAG systems expand an LLM's functionality by connecting it to external databases or specialized modules for tasks like math, law or fact-checking.¹³³ If the LLM recognizes a question requiring technical precision, it can query a module adept at mathematics or legal precedent and then integrate the output into its final text.¹³⁴ This design enables the system to transcend the knowledge limitations inherent in its training corpus.¹³⁵

For governance, RAG adds complexity regarding responsibility. If the final answer is erroneous, is the fault with the LLM's integration logic or the external module's data?¹³⁶ Moreover, sophisticated RAG setups could orchestrate dozens of specialized tools, effectively acting as an executive decision-maker. Policymakers must then decide how to distribute legal obligations across the entire pipeline.¹³⁷

G AI as Field Marshal

I Coordinating Human and Machine Resources

Historically, computers took instructions from humans, but modern AI can direct humans to carry out specific tasks, reversing the dynamic.¹³⁸ In distributed work environments, an LLM could manage schedules, delegate responsibilities and gather feedback from multiple sources.¹³⁹ This phenomenon has already materialized in gig platforms, where algorithmic management sets pay rates, assigns tasks and disciplines workers.¹⁴⁰

Once an AI agent holds such authority, the line between tool and supervisor becomes blurred.¹⁴¹ Some jurisdictions might treat AI as an extension of the employer, holding the employer accountable for biases or errors in algorithmic

131 Chirag Agarwal et al., *Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models*, arXiv 6 (9 February 2024), <https://arxiv.org/abs/2402.04614v3>.

132 Cal. Bus. & Prof. § 17500 (2024) (prohibiting false or misleading advertisements, potentially relevant for AI-based legal advice).

133 Lewis et al., *supra* note 51, at 8-9.

134 *Ibid.*; Weihang Su et al., *DRAGIN: Dynamic Retrieval Augmented Generation Based on the Information Needs of Large Language Models*, arXiv:2403.10081, at 4-9 (21 September 2024), <https://arxiv.org/pdf/2403.10081v3>.

135 Su et al., *supra* note 135, at 8-9.

136 Turner, *supra* note 83, at 170-71.

137 Cohen, *supra* note 36, at 99-101, 170-178.

138 Sayed Fayed Ahmad et al., 'Impact of Artificial Intelligence on Human Loss in Decision Making, Laziness and Safety in Education', 10 *Humanit Soc. Sci. Commun.* 311, at 3 (June 9, 2023).

139 Gray & Suri, *supra* note 111, at 88-90.

140 Imran Kadolkar, Sven Kepes & Mahesh Subramony, 'Algorithmic Management in the Gig Economy: A Systematic Review and Research Integration', *J. Organ. Behav.* (forthcoming), <https://onlinelibrary.wiley.com/doi/full/10.1002/job.2831>.

141 Nell Watson, *Will AI Become Your Assistant or Your Boss? An Ethicist Explains*, Fast Company (11 July 2024), <http://fastcompany.com/91153187/will-ai-become-your-assistant-or-your-boss-an-ethicist-explains>.

decisions.¹⁴² In the past, if the AI was licensed from a third-party vendor, the chain of accountability may have seemed fragmented, suggesting the potential for leaving neither the vendor nor the user with clear responsibility for worker grievances; nevertheless, there have been recent attempts to alleviate this possibility, as regulatory guidance and case law have evolved to accommodate concerns of a fragmented chain of accountability in the event that AI licensed from a third-party vendor might be at issue.¹⁴³

II Implications for Agency and Personhood

The idea of an AI field marshal also challenges the classic legal concept of agency, where a person (the principal) controls an agent who acts on their behalf.¹⁴⁴ If an LLM-based system directs tasks across an organization, does it act as an agent of the developer, the user or itself?¹⁴⁵ Some scholars suggest new legal categories – limited personhood for AI in specific contexts, akin to corporate personhood – to ensure accountability when decisions cause harm.¹⁴⁶

Conversely, critics warn that attributing legal personhood to AI could weaken the claims of injured parties by limiting recourse to the system's owners or developers.¹⁴⁷ As AI's managerial capabilities grow, legislatures and courts will likely face pressure to clarify these issues, building on centuries of precedent surrounding corporations, partnerships and other artificial entities.¹⁴⁸

H Modern AI Governance: Are We Missing Something?

I Incomplete Approaches in Current Frameworks

Despite the proliferation of ethical guidelines, policy toolkits and regulatory proposals, many frameworks fail to account for the lessons gleaned from the hidden history of autonomous agents.¹⁴⁹ For example, risk-based models may attempt to categorize AI systems according to the severity of potential harms but seldom incorporate the possibility of emergent behaviour that transcends the system's initial classification.¹⁵⁰

142 Jon Meredith Garon, *Regulation by the EEOC and the States of Algorithmic Bias in High-Risk Use Cases*, American Bar Association (20 February 2025), https://www.americanbar.org/groups/business_law/resources/business-lawyer/2024-2025-winter/eec-states-regulation-algorithmic-bias-high-risk/.

143 *Ibid.*; Jey Kumarasamy & Brenda Leong, *Third-Party Liability and Product Liability for AI Systems*, IAPP (26 July 2023), <https://iapp.org/news/a/third-party-liability-and-product-liability-for-ai-systems>. Furthermore, as of February 2025, the EEOC has published guidance relevant to antidiscrimination law. And, the IAPP also offers guidance about third-party liability and accountability with respect to AI systems.

144 Restatement (Third) of Agency (Am. Law Inst. 2006).

145 Watson, *supra* note 142.

146 Shlomit Yanisky-Ravid & Xiaoqiong (Jackie) Liu, 'When Artificial Intelligence Systems Produce Inventions: An Alternative Model for Patent Law at the 3A Era', 39 *Cardozo L. Rev.* 2215, 2221-2222 (2018).

147 *Ibid.*, at 2223-2224.

148 *Ibid.*, at 2234-2235.

149 David Leslie & Antonella Maia Perini, 'Future Shock: Generative AI and the International AI Policy and Governance Crisis', Special Issue 5, *Harv. Data Sci. Rev.* (31 May 2024).

150 *Ibid.*

Similarly, transparency mandates often focus on explaining how an AI reaches conclusions but do not address decentralized processes where no single agent has the full picture.¹⁵¹ This gap implies that current governance structures might be too rigid or too narrowly focused on single-agent scenarios, leaving them ill-prepared for multi-agent collectives or orchestrators like LLM field marshals.¹⁵²

II Towards a Layered and Adaptive Governance Model

The complexities of modern AI suggest the necessity of a layered governance structure that incorporates local, regional and even international levels of accountability.¹⁵³ Each layer could address different aspects of autonomy – from algorithmic design standards to real-time monitoring of critical systems – and coordinate with adjacent layers.¹⁵⁴ This approach acknowledges the decentralized nature of multi-agent AI and emergent phenomena.¹⁵⁵

Adaptive regulation also requires iterative review: as AI evolves through machine learning updates, legal frameworks need to be revisited.¹⁵⁶ One strategy involves “sandbox” environments, where innovators can test new AI capabilities under regulatory scrutiny, enabling oversight bodies to update rules before large-scale deployment.¹⁵⁷

III Balancing Innovation and Risk Mitigation

Excessive regulation can stifle beneficial innovation, while too little oversight invites disastrous consequences.¹⁵⁸ Policymakers must aim for an equilibrium that encourages technological progress while safeguarding public welfare.¹⁵⁹ Looking to historical examples – from ancient automata to Shakey – reveals that fear of overreach, liability or unpredictability has always accompanied automated tools.¹⁶⁰

The distinction today is scale and interconnectedness. A flawed mechanical servant in ancient Greece might harm a single household, but a misaligned LLM integrated into financial systems could disrupt global markets.¹⁶¹ This potential amplifies the urgency behind forging robust, historically informed governance measures.¹⁶²

151 Usman Anwar et al., ‘Foundational Challenges in Assuring Alignment and Safety of Large Language Models’, arXiv:2404.09932, at 37-39 (15 April 2024).

152 *Ibid.*

153 *Ibid.*, at 97-101.

154 *Ibid.*, at 97-107.

155 *Ibid.*, at 37-39.

156 *Ibid.*, at 99-100.

157 *Ibid.*, at 36-39.

158 Joe McKendrick, ‘AI Regulation: Too Much Too Soon, or Not Enough?’ *Forbes* (12 November 2024, 3:19 a.m. EST), <https://www.forbes.com/sites/joemckendrick/2024/11/12/ai-regulation-too-much-too-soon-or-not-enough/>.

159 *Ibid.*

160 Renata Włoch, Katarzyna Śledziewska & Satia Rozynek, ‘Who’s Afraid of Automation? Examining Determinants of Fear of Automation in Six European Countries’, 81 *Tech. Soc.* 102782, at 1-2 (June 2025, forthcoming).

161 CFTC & SEC, *supra* note 68, at 2-4.

162 *Ibid.*, at 6-8.

I Conclusion

From the ancient Greek “Automatic Servant” to twenty-first-century LLMs orchestrating tasks across industries, the lineage of autonomous agents offers enduring lessons on control, responsibility and the potential for unexpected outcomes.¹⁶³ Contemporary debates on AI ethics, fairness and accountability often focus on transparency and risk assessment, but they sometimes overlook how decentralized architectures, emergent behaviours and assumptions of inherent beneficence can derail even the best-intentioned regulatory frameworks.¹⁶⁴

The rise of LLM agents, equipped with RAG and capable of directing both humans and machine subcomponents, intensifies these challenges.¹⁶⁵ Traditional command-and-control governance must adapt to a world where AI can serve as a field marshal, coordinating resources and making real-time judgments that humans merely ratify.¹⁶⁶ Emergent forms of multi-agent dynamics, once confined to research laboratories, have become commercial realities in finance, online platforms, healthcare and beyond.¹⁶⁷

By studying this hidden history – cybernetics, early symbolic AI, Braitenberg Vehicles, Boids and other heuristic systems – we unearth critical guidelines for law and policy. Decentralized control implies that no single authority can reliably oversee an entire AI ecosystem. Emergent behaviour demands that the law accommodate the possibility of unforeseen outcomes, calling for agile, layered governance. The assumption of beneficence must be counterbalanced by fail-safes and ethical oversight, while the distinctiveness of LLM-based agents necessitates fresh thinking on liability, personhood and the human-AI relationship.

In short, modern AI governance can benefit substantially from a deeper understanding of past efforts. There are fundamental questions that persist and intensity as AI’s capabilities grow: Who is in control? Who is liable for unforeseen actions? How do we safeguard public welfare without halting innovation? A well-rounded legal structure will incorporate the old and the new, forging a path that recognizes the autonomy and complexity of AI agents while ensuring they are harnessed for the collective good.

163 Mica R. Endsley, ‘From Here to Autonomy: Lessons Learned from Human-Automation Research’, 59 *Hum. Factors* 5 (February 2017).

164 Anwar et al., *supra* note 152, at 37-39, 99-100.

165 *Ibid.*, at 17-21.

166 Ben Green, ‘The Flaws of Policies Requiring Human Oversight of Government Algorithms’, 45 *Comput. L. Sec. Rev.* 105681, at 1-9 (July 2022).

167 Adib Bin Rashid & MD Ashfakul Karim Kausik, ‘AI Revolutionizing Industries Worldwide: A Comprehensive Overview of Its Diverse Applications’, 7 *Hybrid Adv.* 100277, at 10-26 (December 2024).