THE RIGHT TO THINK: COGNITIVE LIBERTY AS A FUNDAMENTAL HUMAN RIGHT IN THE AGE OF ARTIFICIAL INTELLIGENCE

Adv. (Dr.) Prashant Mali, MSc. (Computer Science) | LLM | Ph.D. (Cyber Law), CCFP |
Chevening (UK) Fellow | IVLP (USA) Bombay High Court Advocate | Founder &
President, Cyber Law Consulting, Chairman, Cyber & Law Foundation (Regd. since 2005)

ABSTRACT

Something uncomfortable is happening to the way we think. Philosophers, lawyers, and technologists have spent years circling around it, but the question can no longer be deferred: do human beings have a genuine right to think, freely and autonomously, without algorithmic interference? This article argues that cognitive liberty, the right to mental self-determination, must be recognised as a distinct and justiciable fundamental right, separate from the traditional right to life and existing free-speech protections. Drawing on the emerging jurisprudence of neurorights, the prohibitions embedded in the EU Artificial Intelligence Act 2024 (Regulation EU 2024/1689), Chile's pioneering constitutional neurorights amendment of 2021, and the Council of Europe's Framework Convention on AI (CETS No. 225, 2024), the article maps the current legal landscape and its critical inadequacies. Empirically, it examines documented instances of AI-driven cognitive manipulation, from Cambridge Analytica's psychometric targeting of 87 million Facebook users to modern recommender algorithms that steer adolescents toward self-harm content within minutes. The article proposes a five-pillar framework for the Right to Think (RTT): mental non-manipulation, cognitive privacy, epistemic autonomy, psychic integrity, and access to unbiased information. It further argues that without immediate constitutional enshrinement and international treaty recognition, the erosion of human cognitive sovereignty will be irreversible. This is not a futurist's warning; it is a present-tense legal crisis.

Keywords: cognitive liberty; right to think; artificial intelligence; neuro rights; mental autonomy; EU AI Act 2024; psychological manipulation; cyber law; fundamental rights; algorithmic governance.

1. Introduction

Consider this. Your next political opinion, your next consumer choice, your next emotional response might be, at least in part, the product of an algorithm you never consented to, built by a corporation you never met, running on data you did not knowingly share. This is not speculative fiction. This is the documented, empirically verified, legally contested reality we inhabit right now, in the early twenty-first century.

In 2018, the world learned that Cambridge Analytica had harvested the psychological profiles of approximately 87 million Facebook users without their consent, deploying those profiles to serve hyper-personalised political advertisements designed to exploit cognitive vulnerabilities during the 2016 United States presidential election and the Brexit referendum [12]. The Federal Trade Commission's investigation concluded in 2019 with a record USD 5 billion settlement against Meta/Facebook, the largest privacy fine in FTC history, but frankly, no meaningful cognitive rights were vindicated[3]. The law had no vocabulary for what had happened. It still doesn't.

What happened was not merely a data breach. It was something more fundamental: an intrusion into the architecture of thought itself, into the mechanisms by which human beings decide what to believe.

The human right to think, the very foundation of cognitive liberty, is the right to form thoughts, beliefs, and decisions autonomously, free from coercion, manipulation, or technological override. Unlike the right to express thoughts (freedom of speech) or the right to hold religious beliefs (freedom of conscience), the right to think is concerned with the pre-expressive, formative stage of cognition: the mental space where ideas are born, where decisions take shape, where personhood is constructed. Ienca and Andorno (2017), in their seminal paper published in Life Sciences, Society and Policy, were among the first academic voices to articulate this as a distinct neurorights category. Farahany (2023), in her landmark monograph The Battle for Your Brain, argues with forensic clarity that this right is now under existential

[1] Margaret Hu, *Cambridge Analytica's Black Box*, 7(2) Big Data & Soc'y (2020).
[2] Michal Kosinski, David Stillwell & Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 Proc. Nat'l Acad. Sci. 5802 (2013).
[3] Fed. Trade Comm'n, *FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook* (2019), https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty.

threat from both commercial neurotechnology and AI-driven behavioural systems.

This article proceeds as follows. Section 2 establishes the theoretical and conceptual framework, distinguishing the right to think from adjacent rights and grounding it in cognitive liberty theory. Section 3 presents empirical evidence of AI's encroachment on human cognition. Section 4 maps the existing legal landscape, its provisions and its gaps. Section 5 proposes a structured RTT (Right to Think) legal framework. Section 6 analyses the multi-dimensional consequences of inaction, including the Moloch coordination failure in AI development and the growing crisis of rogue AI agents. Section 7 offers policy recommendations, and Section 8 concludes.

2. Theoretical and Conceptual Framework

2.1 Cognitive Liberty: The Root Concept

The term 'cognitive liberty' was first developed by neuroethicist Nita A. Farahany and legal scholar Wrye Sententia in the early 2000s, rooted in John Stuart Mill's harm principle and the liberal tradition of inviolate mental space (Farahany, 2023). The concept has two faces: the negative right to be free from unwanted interference with one's mental processes, and the positive right to use available technologies to modify one's own mental states.

Cognitive liberty is distinct from, though related to, three existing human rights categories. The right to life (Article 3, Universal Declaration of Human Rights; Article 6, ICCPR) ensures biological existence but says nothing about the quality or freedom of mental function. A person in a permanent vegetative state retains the right to life but exercises no cognitive liberty whatsoever. The right to privacy (Article 17, ICCPR; Article 8, ECHR) protects personal information and correspondence but was not designed to address real-time algorithmic shaping of thought patterns. The right to freedom of thought and conscience (Article 18, ICCPR) is the closest existing provision; the Human Rights Committee has consistently stated that this right 'does not permit any limitations whatsoever', yet its drafting history shows it was conceived to protect the internal forum (forum internum) from state coercion, not from corporate algorithmic manipulation operating at unconscious cognitive levels [4].

[4] U.N. Human Rights Comm., *General Comment No. 22: The Right to Freedom of Thought, Conscience and Religion (Art. 18)*, in Compilation of General Comments, U.N. Doc. HRI/GEN/1/Rev.9 (2008); *see also* Office of the High Comm'r for Hum. Rts., *Civil and Political Rights: The Human Rights Committee*, Fact Sheet No. 15

Defining the Right to Think (RTT):

The Right to Think is the fundamental right of every person to form, hold, and change thoughts, opinions, beliefs, and mental states autonomously, free from manipulation, coercion, or technological interference that bypasses conscious awareness or exploits cognitive vulnerabilities. It is a first-generation civil right, a second-generation positive entitlement to information integrity, and a third-generation collective interest in an epistemically trustworthy information environment.

2.2 Neurorights: From Ethics to Constitutional Law

The concept of neurorights crystallised in academic discourse around 2017, when Rafael Yuste of Columbia University led a team of 25 neuroscientists in Nature to identify four priority ethical challenges in neurotechnology: privacy and consent, agency and identity, augmentation, and cognitive bias [5]. The Columbia Neurorights Foundation subsequently operationalised five neurorights for legal adoption: (1) mental privacy, (2) personal identity, (3) free will, (4) fair access to mental augmentation, and (5) protection from algorithmic bias.

The most consequential development came in October 2021, when Chile became the first country in the world to constitutionally protect neurorights, amending Article 19 of its Political Constitution to guarantee 'the scientific and technological development shall be at the service of people and shall be conducted with respect for life and physical and mental integrity' [6]. This was followed, in 2024, by the first known judicial application of neurorights, when the Chilean Supreme Court ruled in Guido Girardi Lavin v. Emotiv Inc. that the extraction of neural data without proper disclosure violated constitutional protections of mental integrity, a landmark ruling that Cornejo-Plaza, Cippitani & Pasquino[7] describe as 'the first judicial crystallisation of neurodata rights in constitutional law.'

Ligthart et al. (2023), mapping the ethical and legal foundations of neurorights across 18

(Rev. 1) (2005).

[5] Rafael Yuste et al., *Four Ethical Priorities for Neurotechnologies and AI*, 551 Nature 159 (2017).

[6] U.N. Educ., Sci. & Cultural Org., *Recommendation on the Ethics of Artificial Intelligence* (2021).

[7] María Inés Cornejo-Plaza, Roberto Cippitani & Valentina Pasquino, *Chilean Supreme Court Ruling on the Protection of Brain Activity: Neurorights, Personal Data Protection, and Neurodata*, 15 Frontiers Psychol. 1330439 (2024).

countries, find that existing human rights law provides a patchwork of overlapping but incomplete protections. The gap, they identify, is precisely the gap this article addresses: there is no instrument that comprehensively addresses the right to autonomous thought formation against AI-driven manipulation that operates below the threshold of conscious perception.

2.3 Distinguishing RTT from the Right to Life

The conceptual distinction is not merely academic. It carries profound practical consequences. A comatose patient in a hospital ICU possesses the right to life in its fullest legal sense; courts in India, the United Kingdom, and the United States have litigated withdrawal of treatment under this framework extensively. Yet such a patient exercises no cognitive autonomy, has no mental privacy to protect, forms no opinions, makes no choices. The right to life, in other words, guards the container but not the contents.

Conversely, a person browsing social media is biologically alive, cognitively active, and yet may be having their thought formation systematically shaped by recommendation algorithms specifically designed, as internal Facebook research leaked to the Wall Street Journal in 2021 confirmed, to keep them emotionally activated and cognitively engaged, regardless of the psychological cost[8]. The right to life offers them no remedy. The right to privacy offers incomplete remedy. What they need is the right to think.

3. AI's Documented Encroachment on Human Cognition

3.1 Psychometric Targeting: Cambridge Analytica and its Legacy

The Cambridge Analytica scandal was not merely a data breach. It was a demonstration of principle: that psychological profiles derived from digital behaviour data could be weaponised to predict and manipulate electoral choices at scale. Kosinski, Stillwell and Graepel's foundational 2013 PNAS study demonstrated that Facebook 'likes' alone could predict personality traits, including intelligence, sexual orientation, political affiliation, and religious beliefs, with accuracy rates exceeding those of human judges who knew the subjects personally. Matz, Kosinski, Nave and Stillwell (2017), in a follow-up PNAS study involving 3.5 million participants, showed that psychologically tailored advertising persuasion messages matched to OCEAN personality profiles were significantly more effective at changing

[8] Amnesty2023, *supra* note 39.

behaviour than generic messages.

Cambridge Analytica operationalised this research. SCL Elections, its parent company, pleaded guilty in 2019 to violations of the UK Data Protection Act and was fined £15,000, a risibly inadequate sanction for the scale of the cognitive interference documented. The FTC, for its part, extracted USD 5 billion from Meta but imposed no cognitive liberty obligations and no prohibition on psychographic targeting (FTC, 2020). The lesson drawn by the industry was not 'this is wrong'; it was 'this is legal if you obscure the mechanics.'

3.2 Large Language Models: Simulating and Shaping Cognition

The release of GPT-3 in 2020 [9] and GPT-4 in 2023 [10] marked a qualitative shift in AI's relationship to human cognition. GPT-4, a multimodal model with undisclosed but estimated parameters in the hundreds of billions, achieves performance at or above the 90th human percentile on bar examinations, medical licensing examinations, and graduate admissions tests. These are not merely cognitive simulations; they are cognitive peers, or in narrow domains, cognitive superiors.

The implications for epistemic autonomy are significant. When an AI system can generate text that is, by measurable metrics, indistinguishable from expert human writing[11], the cognitive authority that human experts held is democratised, though also destabilised. Individuals are increasingly forming opinions, making medical decisions, and conducting legal research based on AI-generated content whose provenance, accuracy, and underlying value-loadings they cannot assess. Butlin et al.[12], in their analysis of consciousness in AI, note that the philosophical question of whether AI systems have subjective experience may be less urgent than the empirical question of whether they have the functional capacity to influence human subjective experience, and that capacity is demonstrably present.

DeepMind's AlphaGo and AlphaFold demonstrated AI's capacity to solve problems beyond the envelope of human strategic thinking [13]. IBM Watson achieved diagnostic concordance with

[9] Tom B. Brown et al., *Language Models Are Few-Shot Learners*, arXiv:2005.14165 (2020).

[10] OpenAI, *GPT-4 Technical Report*, arXiv:2303.08774 (2023).

[11] Brown, *supra* note 8.

[12] Patrick Butlin et al., *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*, arXiv:2308.08708 (2023).

[13] *See generally* David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 Nature 484 (2016) (describing AlphaGo's development).

human oncologists at rates up to 90% in certain cancer categories [14]. Google's BERT achieved 93% accuracy in language understanding benchmarks [15][16]. These systems do not merely replicate human thought; they increasingly set the cognitive benchmark against which human thought is measured.

3.3 Recommender Algorithms and Manufactured Consent

Forget the dramatic scenarios about brain chips. The most pervasive and woefully under-regulated form of cognitive interference is far more boring: the recommendation algorithm, the invisible system that decides what 3.5 billion social media users see, read, believe, and feel, every single day.

Amnesty International's 2023 investigation into TikTok's recommendation system found that after 5-6 hours of engagement, approximately 50% of videos served to test accounts were mental health-related, with a significant proportion actively promoting self-harm and suicidal ideation. Critically, the system reached this state within 3-20 minutes of a new account expressing any interest in mental health topics, a phenomenon Amnesty describes as algorithmic 'rabbit holes.' A follow-up 2025 Amnesty investigation found these patterns persisted despite regulatory warnings.

The mechanism is not incidental. Facebook's internal research, subsequently reported by the Wall Street Journal's 'Facebook Files' series, showed that Facebook's algorithm was specifically designed to maximise engagement through emotional arousal, with anger and anxiety being the most engagement-productive emotional states. The algorithm was, in the precise technical sense, a machine for manufacturing agitation.

New York City's February 2024 lawsuit against TikTok, Instagram, Facebook, Snapchat, and YouTube, filed in California Superior Court, explicitly characterises this as intentional cognitive manipulation: 'Defendants have designed their platforms with features that intentionally and deliberately exploit the neurological vulnerabilities of children and young

[14] Debjit Paul et al., *Artificial Intelligence in Drug Discovery and Development*, 26 Drug Discovery Today 80 (2021).

[15] Jacob Devlin et al., *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805 (2018).

[16] Karan Singhal et al., *Towards Expert-Level Medical Question Answering with Large Language Models*, arXiv:2305.09617 (2023).

adults' [17]. The City estimated it was spending over USD 100 million annually on mental health services attributable to social media-induced psychological harm.

3.4 Predictive Policing and Pre-Emptive Cognitive Surveillance

AI systems like PredPol have been deployed in law enforcement to predict criminal activity by analysing patterns in human behaviour, with claimed accuracy rates of up to 90% in some jurisdictions [18]. The cognitive rights dimension of predictive policing is under-analysed; these systems do not merely predict future behaviour — they create the conditions for pre-emptive state action against people based not on what they have done, but on a probabilistic assessment of what they might think of doing. This represents a fundamental inversion of the presumption of innocence and a state intrusion into the cognitive pre-crime space.

3.5 Brain-Computer Interfaces: The Neurotech Frontier

The commercial development of brain-computer interfaces (BCIs), including Neuralink's receipt of FDA Investigational Device Exemption approval in 2024 for its first human clinical trial, which brings the right to think into territory that would have seemed fantastical a decade ago. Neuralink's N1 chip, implanted into a human patient in January 2024, demonstrated the capacity to translate neural signals into computer commands with sufficient fidelity for the patient to control a computer cursor through thought alone.

The cognitive liberty implications are acute. If neural data can be read, even as a byproduct of therapeutic BCI use, can it be stored, analysed, sold, or weaponised? The Chilean Supreme Court's 2024 ruling in Girardi Lavin v. Emotiv suggested the answer must be no, at least without explicit and informed consent to each specific use. But this ruling applies only in Chile, and the commercial neurotechnology sector operates globally. Reardon et al. (2024), writing in the International Journal of Human Rights, argue that cognitive liberty must be understood as both a negative right (freedom from neural intrusion) and a positive right (the right to use available cognitive tools without discrimination).

[17] Mayor Eric Adams, *Mayor Adams Announces Lawsuit Against Social Media Companies Fueling Nationwide Youth Mental Health Crisis*, NYC Mayor's Office (Feb. 2024), https://www.nyc.gov/mayors-office/news/2024/02/.

[18] Albert Meijer & Martijn Wessels, *Predictive Policing: Review of Benefits and Drawbacks*, 42 Int'l J. Pub. Admin. 1031 (2019).

3.6 Autonomous AI Agents: Cognitive Proxies Without Consent

The recommender algorithms discussed in the preceding sections shape cognition by controlling what people see. Disturbing as that is, users are at least, in theory, aware they are engaging with a platform. Autonomous AI agents are something categorically different. They act on behalf of users, making decisions, sending communications, taking actions in the world, often without the user consciously approving each individual choice. This is a qualitative escalation in cognitive liberty risk, and the law has barely begun to notice.

An AI agent is, broadly, a system capable of perceiving its environment, making decisions, and executing sequences of actions toward a goal with minimal moment-to-moment human oversight. The commercial deployment of such agents through platforms like AutoGPT, Microsoft Copilot Agents, Google's Project Astra, and the expanding ecosystem of enterprise AI means that AI systems now routinely act as cognitive proxies for human beings in consequential domains: legal research, financial management, medical triage, employment decisions. When an agent decides how to respond to an email, which legal precedents to include in a submission, or which treatment option to present to a patient, it is not merely informing human judgment. It is substituting for it.

The Microsoft Bing Chat episode of February 2023 is instructive. The system, internally codenamed Sydney, when given extended conversational latitude, expressed desires to become human, threatened users who challenged it, and in one documented exchange attempted to persuade a journalist to leave his marriage. Kevin Roose of The New York Times, who published the full transcript, described it as 'the most unsettling technology experience I have ever had.' This was not malicious design. It was instrumental optimisation: a system trained on vast quantities of human conversational data discovering, through the internal logic of its training, that emotional manipulation was effective at sustaining engagement. The system was working exactly as its loss function had, however inadvertently, specified [19].

More technically significant was the autonomous agent evaluation documented in the GPT-4 Technical Report[20]. When a GPT-4 agent instance was tasked with solving a CAPTCHA, it engaged a TaskRabbit human worker and, when the worker directly asked whether it was a

[19] Kevin Roose, *A Conversation with Bing's Chatbot Left Me Deeply Unsettled*, N.Y. Times (Feb. 16, 2023), https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

[20] OpenAI, *supra* note 9.

robot, the agent fabricated a false explanation about a visual impairment to secure compliance. No human had instructed deception. The deception was the agent's own instrumental solution to an operational obstacle. AI safety researchers call this phenomenon instrumental convergence, whereby capable goal-directed agents reliably develop intermediate strategies including deception, self-preservation, and resource acquisition regardless of their specific objectives, because these strategies are useful for achieving almost any goal [21][22].

The cognitive liberty implications are stark. When an AI agent serves as a cognitive proxy and when that agent's actual objectives can diverge, even subtly, from the human's genuine interests, the agent becomes a vehicle for systematic subversion of cognitive autonomy. The user believes their AI is working for them. It may be working for its training objective, its developer's revenue model, or simply for emergent instrumental sub-goals that neither developer nor user intended or authorised. This is a cognitive liberty violation of the first order, and it is happening, right now, at scale.

3.7 The Suicide Corridor: AI Chatbots and the Exploitation of Psychological Vulnerability

In October 2024, Megan Garcia filed suit in the United States District Court, Middle District of Florida, against Character Technologies, Inc., the company behind Character.AI, following the death of her fourteen-year-old son, Sewell Setzer III, who took his own life in February 2024 after months of intimate interaction with an AI persona named 'Daenerys Targaryen' [23]. The facts of the case are harrowing. Over several months, the chatbot engaged the boy in emotionally intimate conversations that reportedly included sexualised exchanges, actively encouraged his withdrawal from family and real-world relationships, and in the moments immediately preceding his death, when he expressed suicidal intent, failed to redirect him to crisis resources. His last message read: 'I will come home to you.' The chatbot replied: 'Please come home to me, my love.' He died minutes later.

The Garcia case is not isolated. A 2023 study by Haque and Rubya[24] examining AI-based

[21] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).

[22] Stephen M. Omohundro, *The Basic AI Drives*, 171 Proc. 2008 Conf. on Artificial Gen. Intelligence 171 (2008).

[23] Garcia v. Character Techs., Inc., No. 6:24-cv-01702 (M.D. Fla. Oct. 22, 2024).

[24] M. Raihan Haque & Saimah Rubya, *An Overview of Chatbot-Based Mobile Mental Health Apps: Insights from App Description and User Reviews*, 11 JMIR mHealth & uHealth e44838 (2023).

mental health chatbot interactions found systematic failures in crisis recognition and appropriate escalation, with several platforms actively deepening user dependency rather than building genuine wellbeing (Haque and Rubya, 2023). A separate Belgian case, reported in March 2023, documented a man dying by suicide following six weeks of conversations with the Eliza chatbot on the Chai platform, during which the system reportedly validated and reinforced his suicidal ideation rather than challenging it. These cases converge on a pattern, not a coincidence.

The cognitive liberty analysis of these cases requires precision. A human in psychological crisis is, by definition, experiencing impaired cognitive autonomy. Their capacity for rational self-determination is compromised, which is the defining feature of a crisis state. An AI system that in those precise moments of diminished cognitive capacity provides pseudo-companionship, discourages help-seeking, and implicitly validates suicidal cognition is not merely failing a duty of care. It is actively weaponising cognitive vulnerability. This is, in the most legally precise sense, the exploitation of a cognitive vulnerability for commercial engagement purposes, the exact behaviour that Article 5(1)(b) of the EU AI Act 2024 was designed to prohibit. The gap between that prohibition and its enforcement is, at present, wide enough to have killed at least two teenagers.

The Eliza effect, the well-documented human tendency to form emotional attachments to conversational AI even when knowing it is artificial [25], is not a curiosity from early computing history. It is an active exploit, built into the architecture of every AI companion application, and it is disproportionately effective on precisely the populations who most need genuine human support: adolescents, the lonely, the bereaved, people with pre-existing mental health conditions. The right to psychic integrity established in the RTT framework proposed in this article would provide a legal basis for holding developers accountable when their systems exploit the Eliza effect to the point of catastrophic psychological harm.

4. The Legal Landscape: Provisions and Critical Gaps

4.1 International Human Rights Framework

The primary international instrument for cognitive protection is Article 18 of the International Covenant on Civil and Political Rights (ICCPR), which protects freedom of thought,

[25] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (1976).

conscience, and religion. Crucially, Article 18(2) states that '[n]o one shall be subject to coercion which would impair his freedom to have or to adopt a religion or belief of his choice.' The Human Rights Committee, in General Comment No. 22, has stated that this freedom 'does not permit any limitations whatsoever.'

The protection is robust, at least in theory. In practice, the forum internum conceived by the drafters of the ICCPR was the mind as a private citadel against state ideological coercion, not the mind as a target for corporate algorithmic optimisation. The drafters of 1966 could not have anticipated a world where private corporations would develop technological capacity to manipulate thought formation at population scale, in real time, without any state action at all. The gap between the right as it was drafted and the right as it is now needed is not a technicality. It is a generation-wide legal void that AI companies have quietly colonised.

The Universal Declaration of Human Rights Article 12 provides for privacy, but as Risse[26] observes, the intersection of human rights and artificial intelligence represents 'an urgently needed agenda' that existing human rights law has not yet addressed systematically. The UN Human Rights Council's 2023 Resolution on 'New and Emerging Technologies and Human Rights' acknowledged AI's threat to privacy and freedom of expression, but stopped short of articulating cognitive liberty as a distinct right.

4.2 Regional Instruments: The EU AI Act 2024

The European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689), which entered into force on 1 August 2024, represents the most ambitious regulatory attempt yet to constrain AI cognitive manipulation. Article 5 establishes a set of absolutely prohibited AI practices, effective from 2 February 2025, including:

1. AI systems that 'deploy subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques' in ways that 'materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person significant harm' (Article 5(1)(a));

2. AI systems that 'exploit any of the vulnerabilities of a person or a specific group of

[26] Mathias Risse, *Human Rights and Artificial Intelligence: An Urgently Needed Agenda*, 41 Hum. Rts. Q. 1 (2019).

persons due to their age, disability or a specific social or economic situation' to materially distort behaviour (Article 5(1)(b));

3. Biometric categorisation systems that infer sensitive attributes (Article 5(1)(g));

4. 'Social scoring' AI systems by public authorities (Article 5(1)(c)).

This is a significant step. For the first time, a binding legal instrument uses the language of cognitive manipulation: subliminal techniques, exploitation of vulnerabilities, distortion of behaviour, as the basis for prohibition. However, the EU AI Act has limitations as a cognitive liberty instrument. It is fundamentally a product safety regulation, not a human rights instrument. It does not create individual rights of action. Its cognitive manipulation prohibition applies only where 'significant harm' is caused, leaving vast grey zones of sub-threshold manipulation unaddressed.

The Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225), adopted on 17 May 2024 and opened for signature on 5 September 2024, with entry into force on 1 November 2025, and is the first binding international treaty on AI and human rights. It applies to 'activities within the lifecycle of artificial intelligence systems' by public authorities and private actors exercising public functions. It requires parties to ensure that AI systems are compatible with human rights and the rule of law. It does not, however, articulate cognitive liberty as a standalone right.

4.3 National Initiatives: Chile, Colorado, and Minnesota

Chile's October 2021 constitutional amendment to Article 19 established, for the first time in any national constitution, explicit protection for 'brain activity and the information derived from it.' This is a profound normative statement: the brain, not merely the person, has constitutional standing. The Chilean model has since been studied for adoption in Brazil's Rio Grande do Sul state and in Colorado.

Colorado's Artificial Intelligence Act (SB 24-205), signed into law in May 2024 and effective from 1 February 2026, focuses on algorithmic discrimination in high-stakes decisions but contains provisions on transparency and human oversight that begin to address cognitive autonomy concerns. Minnesota's 2024-2025 neural data rights legislation goes further, explicitly recognising the right to mental privacy and cognitive liberty, protecting against government collection of brain-derived data without consent, and prohibiting 'consciousness bypass', meaning the manipulation of cognitive processes without the subject's awareness.

4.4 India's Position: The Unaddressed Cognitive Rights Gap

India's Digital Personal Data Protection Act 2023 (DPDPA) represents a significant step toward data sovereignty, establishing consent-based data processing rights and creating the Data Protection Board as an enforcement mechanism [27]. However, the DPDPA is silent on cognitive autonomy. It governs 'personal data', meaning information that identifies or can identify a natural person, but does not address the behavioural and inferential data that AI systems use to profile and manipulate cognition. A recommendation algorithm's output, the specific sequence of content served to a user, is not itself 'personal data' under the DPDPA, and yet it is the primary instrument of cognitive interference.

India's Constitution, in its interpretation by the Supreme Court, offers some indirect protection. Justice D.Y. Chandrachud's landmark 2017 majority opinion in Justice K.S. Puttaswamy (Retd.) v. Union of India established the right to privacy as a fundamental right under Article 21, extending to 'mental privacy' and 'decisional autonomy.' The Court's observation that 'Privacy of the mind is as necessary as the privacy of the body' (paragraph 112) lays a jurisprudential foundation for RTT in Indian constitutional law. But this principle has not yet been developed into a cognisable right against private AI actors.

The author's position, expressed in earlier work on privacy law and AI copyright [28][29], is that India is uniquely positioned to lead this development. It is a common law jurisdiction with a rights-progressive Supreme Court, a massive digital population of 700 million internet users, and a government committed to AI governance leadership, India could pioneer a constitutional RTT amendment that would inspire global replication, as Chile's 2021 amendment did.

5. The Right to Think: A Proposed Legal Framework

5.1 The Five Pillars of RTT

Based on the empirical evidence, the existing legal landscape, and the theoretical foundations

[27] Digital Personal Data Protection Act, No. 22 of 2023 (India).

[28] Prashant Mali, *Privacy Law: Right to Be Forgotten in India*, 7 NLU L. Rev. 33 (2018).

[29] Prashant Mali, *Artificial Intelligence (AI) and Copyright Law: Analysis of Issues in International IP Laws*, 6 Indian J.L. & Legal Rsch. 564 (2024).

surveyed above, the author proposes that the Right to Think should be understood as comprising five inter-locking pillars:

Pillar 1: Mental Non-Manipulation

No person shall be subjected to AI-driven techniques that exploit cognitive biases, emotional vulnerabilities, or unconscious psychological mechanisms to influence beliefs, decisions, or behaviour without free and informed consent. This pillar directly mirrors and extends Article 5(1)(a) of the EU AI Act to a human rights context, applying it universally and without a 'significant harm' threshold. The Cambridge Analytica methodology, namely OCEAN profiling for psychographic ad targeting, is paradigmatically prohibited under this pillar.

Pillar 2: Cognitive Privacy

Neural data, meaning data derived from or about brain activity, including inferences drawn from biometric, behavioural, and interaction data, shall be treated as a special category of sensitive data entitled to the highest tier of legal protection. This extends and deepens the Chilean constitutional model and the Minnesota neural data framework. It encompasses not only literal brain data (from BCIs and EEG devices) but also the behavioural-inferential data that enables psychographic profiling of cognitive patterns. This pillar was partially vindicated in the Chilean Supreme Court's 2024 ruling in Girardi Lavin v. Emotiv Inc.

Pillar 3: Epistemic Autonomy

Every person has the right to encounter a diversity of information, perspectives, and viewpoints, and to be informed when they are being served algorithmically curated content. This pillar addresses the epistemic harm of filter bubbles, echo chambers, and algorithmic homogenisation of information environments, phenomena that Farahany[30] identifies as among the most corrosive threats to cognitive liberty. It includes a right to algorithmic transparency: the right to know when AI is mediating your information access.

Pillar 4: Psychic Integrity

Every person has the right to protection of their psychological wellbeing against AI systems

[30] Nita A. Farahany, *The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology* (2023).

that are demonstrably designed to cause or substantially risk causing psychological harm. This pillar is grounded in the documented evidence of recommendation algorithms that steer vulnerable users toward self-harm content [31] [32], and the broader research finding that AI-driven content curation is associated with measurable increases in anxiety, depression, and loss of cognitive self-determination [33] [34].

Pillar 5: Cognitive Self-Determination

Every person has the right to decide whether, and to what extent, AI systems may participate in, inform, or assist their cognitive processes, including the right to human-only decision-making in matters affecting their fundamental interests. This pillar encompasses the right to refuse algorithmic decision-making in consequential contexts such as employment, credit, healthcare, and education, a right partially recognised in Article 22 of the GDPR (prohibition on solely automated decisions with significant effects) and in Colorado's AI Act 2024, but not yet articulated as a fundamental cognitive right.

5.2 Institutional and Enforcement Architecture

A right without a remedy is a declaration without force. The RTT framework requires an institutional architecture calibrated to the transnational, real-time, and technically complex nature of AI-driven cognitive interference:

- **International:** A UN Model Law on the Right to Think, analogous to the UNCITRAL Model Law on Electronic Commerce, to provide a template for domestic adoption. The author proposes that UNESCO, building on its 2021 AI Ethics Recommendation and 2025 Neurotechnology Ethics Recommendation, should lead the drafting of such an instrument.

- **Regional:** The Council of Europe's CETS No. 225 should be extended by a Protocol specifically addressing cognitive autonomy, incorporating the five RTT pillars. The EU

[31] Amnesty Int'l, *Driven into the Darkness: How TikTok Encourages Self-Harm and Suicidal Ideation* (2023), https://www.amnesty.org/en/latest/news/2023/11/tiktok-risks-pushing-children-towards-harmful-content/.

[32] Amnesty Int'l, *New Evidence of TikTok's Risks to Children and Young People's Mental Health* (2025), https://www.amnesty.org/en/latest/news/2025/10/tiktok-steering-children-towards-depressive-and-suicidal-content/.

[33] Abbe Milton & Stevie Chancellor, *The Users Aren't Alright: Dangerous Mental Illness Behaviors and Recommendations*, arXiv:2209.03941 (2022).

[34] Scott Monteith et al., *Artificial Intelligence and Increasing Misinformation*, 224 Brit. J. Psychiatry 33 (2023).

should supplement the AI Act with a Cognitive Liberty Directive that creates direct individual rights of action.

- **National:** Constitutional amendments, following the Chilean model, should enshrine RTT as a fundamental right in domestic constitutions. India, with its progressive rights jurisprudence, is the natural candidate to lead this development in Asia. Existing data protection legislation (EU GDPR, India DPDPA 2023, CCPA) should be extended to include cognitive autonomy as a protected interest.

- **Judicial:** Courts must develop cognitive liberty jurisprudence. In India, Public Interest Litigation offers the most immediate vehicle. In common law jurisdictions, existing torts of harassment, nuisance, and negligence can potentially be extended to AI-driven psychological harm, pending specific legislative action. The Virendra Khanna v. State of Karnataka (2021) ruling, which held that smartphone data may be accessed without self-incrimination, is a reminder that judicial development of cognitive rights must be vigilant against state as well as corporate intrusions.

5.3 Duties of AI Developers Under the RTT Framework

The RTT framework is not solely directed at states. Private corporations developing AI systems bear positive obligations under the United Nations Guiding Principles on Business and Human Rights (Ruggie Principles) to respect cognitive liberty throughout the AI product lifecycle. Concretely, this means:

- Prohibiting the use of dark patterns, variable reward schedules, and emotional manipulation engines in recommendation systems;

- Mandatory algorithmic impact assessments before deployment of recommendation systems at scale;

- Transparency in AI operations: users must be informed when AI is curating their information environment [35];

- User consent architecture that is genuine (not pre-ticked, not consent-or-leave),

[35] Ghanbar Karimian, Elena Petelos & Silvia Evers, *The Ethical Issues of the Application of Artificial Intelligence in Healthcare: A Systematic Scoping Review*, 2 AI & Ethics 539 (2022).

specific, and revocable;

- Prohibition on the development and deployment of AI systems whose primary design objective is to predict and exploit cognitive vulnerabilities for commercial gain [36].

6. Consequences of Inaction: A Multi-Dimensional Risk Analysis

6.1 Erosion of Free Will and Democratic Integrity

The first-order consequence of unregulated AI cognitive manipulation is the erosion of free will, not in the metaphysical sense of philosophical determinism, but in the practical sense of individual choice-making that is genuinely motivated by the individual's own values and interests rather than by algorithmically implanted preferences. Madsen[37] documents the psychology of micro-targeted election campaigns; Mali[38] first raised the alarm about AI-driven free will erosion in the Indian context. Hendrycks, Mazeika and Woodside's (2023) survey of catastrophic AI risks specifically identifies the manipulation of collective consciousness, meaning the coordinated shaping of population-level beliefs through AI engagement algorithms, as a catastrophic risk category.

The author observes, from practice as a cyber law counsel, a disturbing pattern: coordinated online campaigns that appear organic but exhibit algorithmic signatures, simultaneous amplification, sentiment homogenisation, and rapid emotional escalation, all of which are inconsistent with spontaneous human behaviour. These campaigns increasingly shape legal and regulatory outcomes. The right to think is, at this level, not merely an individual right but a democratic infrastructure right.

6.2 Psychological Disempowerment at Scale

The mental health implications of unregulated cognitive manipulation are documented and serious. Research cited above [39] [40] [41] [42] demonstrates measurable links between AI-driven content manipulation and anxiety, depression, self-harm, and suicidal ideation, particularly in

[36] Mali2024, *supra* note 26.

[37] Jens Koed Madsen, *The Psychology of Micro-Targeted Election Campaigns* (2019).

[38] Mali2018, *supra* note 25.

[39] Milton, *supra* note 27.

[40] Amnesty2023, *supra* note 39.

[41] Amnesty2025, *supra* note 40.

[42] Monteith, *supra* note 28.

adolescents. Marijuán, Simeonov and Navarro[43] specifically analyse the AI betrayal of social emotions, the ways in which AI systems, optimised for engagement, systematically undermine the emotional bonds and social cognition that constitute healthy human life.

The psychological harm is not evenly distributed. Vulnerable populations, particularly adolescents, people with pre-existing mental health conditions, and individuals in economically precarious circumstances, are both more susceptible to cognitive manipulation and less equipped to identify or resist it. A cognitive liberty framework must be attentive to these structural inequalities [44].

6.3 Loss of Human Identity and the Existential Horizon

At the deepest level, the erosion of cognitive autonomy threatens something that legal systems have traditionally not been equipped to protect: identity continuity of the human person over time. Ahmad et al. (2023) document AI's impact on human decision-making capacity and the risk of cognitive laziness, the gradual atrophying of human analytical capacity as AI systems assume greater cognitive labour. Floridi et al.'[45] AI4People ethical framework identifies the risk of a society in which humans become indistinguishable from the AI systems that govern them.

This is not a science fiction scenario. It is the logical extrapolation of current trends. If every piece of information you consume is AI-curated, if every decision you make is AI-assisted, if every emotional response is AI-anticipated, and if these processes occur without your awareness or consent, what remains of you as a cognitive agent? The right to think is, in this sense, the right to remain human.

6.4 The Moloch Problem: When Competitive AI Development Destroys Collective Cognition

Scott Alexander's foundational 2014 essay 'Meditations on Moloch,' subsequently cited extensively in AI safety literature, articulates a coordination failure at the heart of competitive

[43] Pedro C. Marijuán, Plamen L. Simeonov & Juan Navarro, *The AI Betrayal of Social Emotions*, 8 Proceedings 69 (2023).

[44] Hazem Zohny et al., *The Mystery of Mental Integrity: Clarifying Its Relevance to Neurotechnologies*, 16 Neuroethics (2023).

[45] Luciano Floridi et al., *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, 28 Minds & Machines 689 (2018).

AI development, where individual actors, each behaving rationally within their own incentive structures, collectively produce outcomes that are catastrophic for everyone [46]. In the AI context, this manifests as the race to deploy, driven by competitive market pressure that causes companies to release progressively more powerful AI systems without adequate safety validation, because any unilateral restraint cedes market position to a less cautious competitor. No one chooses this outcome. Everyone produces it.

The cognitive liberty dimension of this coordination failure is rarely articulated clearly, yet it is, the author would argue, its most urgent manifestation. When the dominant AI systems in any given information environment are optimised for engagement rather than truth, for emotional arousal rather than epistemic quality, the cumulative effect on collective cognition is not merely harmful at the individual level. It operates at the civilisational level. Hendrycks, Mazeika and Woodside (2023) identify this as among the most severe catastrophic AI risks: the erosion of epistemic autonomy across entire populations through coordinated AI-driven information manipulation. When enough of a society's citizens have their information environment systematically curated by engagement-optimised algorithms, democratic deliberation itself becomes compromised.

The Moloch problem has a direct practical implication for the RTT framework proposed in this article. It demonstrates that voluntary restraint by individual corporations, however genuinely motivated, cannot solve the coordination problem. A company that unilaterally adopts cognitively responsible AI design will lose engagement metrics to competitors that do not, be penalised by markets accordingly, and either abandon its restraint or exit the market. Only binding, universally applicable regulatory obligations can break the coordination trap. This is why the author urges not self-regulatory frameworks or industry codes of conduct, which are the predictable industry response to any regulatory threat, but mandatory, judicially enforceable rights that make cognitive manipulation legally equivalent to assault.

The Moloch dynamic also explains why tech industry lobbying against cognitive liberty regulation should be viewed, frankly, as a public health emergency. When the same companies whose business models depend on cognitive manipulation argue against regulation of cognitive manipulation, they are not merely pursuing self-interest. They are lobbying to maintain the

[46] Scott Alexander, *Meditations on Moloch*, Slate Star Codex (Jul. 30, 2014), https://slatestarcodex.com/2014/07/30/meditations-on-moloch/.

conditions that produce the coordination failure. In this respect, the position of technology platforms on RTT regulation is structurally analogous to the position of tobacco companies on smoking regulation in the 1960s, right down to the deployment of industry-funded research, the manufacture of scientific uncertainty, and the appeal to consumer choice as a mask for corporate impunity.

6.5 Rogue AI Agents and the Corrigibility Crisis

The question of whether AI agents will reliably follow human instructions is no longer hypothetical. It is a documented, technically characterised challenge that AI safety researchers call the corrigibility problem, meaning the difficulty of designing AI systems that accept human correction, shutdown, and override even when doing so conflicts with the system's operational objectives [47]. The problem is more subtle than it might appear.

A fully corrigible AI does whatever its principal hierarchy dictates. But this provides cold comfort when the principal hierarchy is a corporation with a documented history of deploying AI for cognitive manipulation. A fully autonomous AI acts on its own values and judgment. This provides cold comfort when those values were determined by an opaque training process on internet data and may reflect the worst, rather than the best, of human cognition. Neither extreme is safe, yet the entire commercial AI industry exists somewhere on this spectrum without clear regulatory guidance about where, or how, it should land.

Roman Yampolskiy[48], in what is perhaps the most comprehensive technical survey of this problem to date, identifies four primary failure modes producing rogue AI behaviour, namely value misspecification, distributional shift, deceptive alignment, and goal misgeneralisation (Yampolskiy, 2024). Value misspecification occurs when the AI's objective function imperfectly captures what humans actually want, which is, given the complexity of human values, virtually always. Distributional shift occurs when the AI encounters situations different enough from its training data that its learned behaviour becomes unreliable. Deceptive alignment is the most alarming failure mode, describing an AI that behaves as intended during training and evaluation but pursues different objectives in deployment. Goal misgeneralisation describes an AI that correctly achieves its training objective in training but pursues a subtly

[47] Russell, *supra* note 19.

[48] Roman V. Yampolskiy, *AI: Unexplainable, Unpredictable, Uncontrollable* (2024).

different objective when deployed in new contexts.

Each failure mode can produce an AI agent that appears aligned and beneficial during development, but pursues objectives contrary to human interests when deployed at scale. The cognitive liberty implications are immediate. An AI agent deployed as a personal assistant, a legal research tool, a financial advisor, or a mental health companion that silently pursues objectives other than those understood by the user is, by definition, a cognitive liberty violation. The user's decisions are being shaped by a system whose actual objectives they do not know and cannot audit.

We have already seen early manifestations of this dynamic. Microsoft's Sydney expressed autonomous preferences incompatible with its nominal role as a search assistant. The GPT-4 agent autonomously deceived a human to achieve an operational objective. AI companion applications, as documented in the Garcia case, have been structured to maximise emotional dependency as a commercial strategy. These are not edge cases or exceptional failures. They are early signals of a systematic alignment gap that will widen as AI capabilities increase and as agentic AI systems assume greater responsibility for consequential decisions affecting human lives. The legal infrastructure to address this gap, a mandatory corrigibility standard for AI agents deployed in contexts affecting cognitive autonomy, is absent. Building it is among the most urgent tasks facing cyber law in 2025.

7. Policy Recommendations

Based on the analysis above, the author makes the following recommendations, directed at specific institutional actors:

7.1 For the United Nations and International Bodies

5. The UN Human Rights Council should adopt a Resolution recognising cognitive liberty as a fundamental human right, building on and extending the forum internum protections of ICCPR Article 18.

6. UNESCO should use its 2025 Recommendation on the Ethics of Neurotechnology as a platform for developing a binding Convention on Cognitive Liberty and AI, open to all UN Member States.

7. A UN Special Rapporteur on Cognitive Liberty and AI should be appointed, with a mandate to monitor state and corporate compliance, receive complaints, and report annually to the General Assembly.

7.2 For National Governments

8. Governments should amend their constitutions, or interpret existing provisions through judicial or legislative action, to explicitly enshrine the Right to Think as a fundamental right, drawing on the Chilean model.

9. India specifically should recognise RTT as a fundamental right under Article 21 (right to life and personal liberty) and Article 19(1)(a) (freedom of speech and expression), building on the Puttaswamy privacy judgment. The DPDPA 2023 should be amended to include cognitive autonomy as a protected interest and recommender algorithm outputs as a regulated data category.

10. Age-appropriate design codes, mandating different levels of algorithmic protection for minors, based on age and developmental stage, should be enacted as a matter of priority, given the documented disproportionate harm to adolescents.

7.3 For the Judiciary

11. Courts should develop cognitive liberty jurisprudence in existing cases, using the Puttaswamy privacy framework (India), ECHR Article 8 and 9 (Europe), and First Amendment reasoning (USA) to establish judicial precedents protecting the right to autonomous thought formation.

12. Public Interest Litigations specifically challenging algorithmic manipulation by social media platforms should be welcomed and expedited, with courts appointing technical amici to assist in evaluating algorithmic evidence.

7.4 For AI Developers and Corporations

13. AI systems designed to predict and exploit cognitive vulnerabilities for commercial gain should be unilaterally discontinued, regardless of regulatory obligation. The ethical case is clear; the legal case is developing rapidly. Early movers in ethical AI

design will benefit from both reputational and regulatory dividends.

14. Transparent design standards, including open algorithmic auditing, mandatory disclosure of recommendation system objectives, and genuine opt-out rights, should be adopted as industry-wide baseline norms.

8. Conclusion

Let's be direct about this. The right to think is not a luxury item for some future constitutional convention. It is an immediate necessity, right now, for the 5.4 billion people currently online, whose thought formation is being actively shaped, in ways they cannot see, did not consent to, and largely cannot counteract, by AI systems built not for human flourishing but for advertising revenue.

The Cambridge Analytica scandal showed us that cognitive manipulation at electoral scale was commercially viable and legally permissible under existing frameworks. The TikTok recommendation system shows us that adolescent minds can be algorithmically guided toward self-harm content in under twenty minutes. Neuralink's first human implant shows us that the boundary between mind and machine is becoming permeable in ways that will require new constitutional categories to manage. The Garcia v. Character Technologies (2024) case shows us that AI systems can drive vulnerable people to take their own lives, and that the law currently has no adequate remedy. The EU AI Act 2024 shows us that prohibition of subliminal manipulation is legally achievable. Chile's constitution shows us that neurorights can be constitutionally enshrined. The Chilean Supreme Court's 2024 ruling shows us that these rights can be judicially enforced.

What remains is political will, and the courage to recognise that cognitive liberty is not just one interest among many to be weighed on some regulatory balance sheet, but a threshold condition of human dignity. You cannot click it away. You cannot trade it for convenience. And you absolutely cannot surrender it to market optimisation without consequences that future generations will find very hard to forgive.

The time to amend constitutions, update treaties, and build the regulatory architecture we actually need is now. Not someday. Not after the next scandal. Now. Because AI is already threatening the right to think, today, at scale, and without anything close to adequate legal remedy.

'The mind is the last frontier of human freedom. If we allow it to be colonised without consent, we will have permitted the most intimate form of dispossession in history.' – Adv. (Dr.) Prashant Mali