Reconfiguring the social turn of multi-agent systems research in the age of agentic genAI

Anonymous Author(s)

ABSTRACT

Since the release into the wild of ChatGPT in late 2022, advances in LLMs have enabled an exponential growth in AI agent-building capacity. The desired results range from increased efficiency to solving complex e.g. climate change related societal problems. Yet, there is also an increase of evil agents that are deliberately resistant to safety training techniques. In the middle ground, increase of agents – even if not deliberately malignant – causes systemic problems such as human redundancy, reinforced systemic biases. As the main resources for AI agent development concentrate in the hands of the Big Tech, **what are the systemic levers to be applied so that we can ensure that the behavior of AI agents aligns with human values and societal goals**. Our proposed solution, the AiCrit tool, proposes a safety by architecture concept that runs deeper than safety by design.

KEYWORDS

LLM, Agentic AI, AI-human alignment, AI and the law

ACM Reference Format:

Anonymous Author(s). 2026. Reconfiguring the social turn of multi-agent systems research in the age of agentic genAI. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 4 pages.

1 INTRODUCTION

The current trajectory of Multi-Agent Systems (MAS) is defined by an exponential growth in autonomy and complexity. However, this development is largely concentrated within Big Tech silos, creating a Black Box effect that disenfranchises the indirect stakeholders such as citizens who are not active developers or users but whose lives are fundamentally influenced by systemic AI biases and emergent deceptive behaviors. Current safety paradigms, such as Reinforcement Learning from Human Feedback (RLHF), are increasingly bypassed by misaligned systems or diluted by technical objectives that prioritize efficiency over societal alignment. General debates on ethics of and for AI and value alignment are mainly focussing on the abstract level of all AI complying with societal or even universal values, resulting in decontextualized solutions that fail when meeting the real world outside of the AI lab and its benchmarking test-sets.

We propose a shift toward a social-contextual focus, where AI agents are designed to protect those who lack a direct voice in the development process. We connect this to the high-level intentions of the EU Digital Services Act (DSA) and the AI Act, interpreting them not merely as technical compliance checklists, but as mandates for a proactive Democracy Shield. Societal safeguards should be implemented at the architectural level rather than relying on bypassed alignment techniques like RLHF. As a concrete realization of this vision, we present AiCrit, an agentic framework implementing Safety-by-Architecture (SbA). Unlike Big Tech-dominated black-box agents that amplify systemic biases or deceptive behaviours, AiCrit enforces governance through foundational invariants: topological constraints prevent infinite loops of self-justifying misalignment, formal policy registries (via JSON Schema) translate abstract legal red lines (e.g. EU Charter protections on expression, misinformation risks) into verifiable requirements, and a coordinator agent acts as a citizen proxy to identify overlooked normative boundaries. This shifts MAS from efficiency-driven silos to socially embedded systems that proactively shield indirect stakeholders, diversifying alignment beyond developer-user dyads and enabling contextualized, human-override moral judgments in real-world deployment.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.

2 THE EU LEGAL PERSPECTIVE

From a human-centric point of view, it must be said that the idea of artificial intelligence (AI) assessing and determining what humans can and cannot communicate via online platforms is challenging. Already, the possibility for natural and legal persons to restrict the freedom of expression of others is limited under Article 11(1) of the Charter of Fundamental Rights of the European Union (EU Charter). Few Europeans would happily give up their privacy and accept unlimited wiretapping and interception of phone calls even for the purpose of avoiding violation of the rights or interests of others. In addition, AI-powered monitoring would arguably subordinate human nature to machine standards as the tools for inter-human communication would effectively confine the space for human expression. Moreover, a platform account provides a more powerful means for mass communication than a telephone subscription, and AI facilitates distortion of information and dissemination of misinformation. A case in point is the AI system called Grok provided by X, which enables users to virtually undress people, including children, and share AI generated adult imagery on the general X platforms. From a law enforcement perspective, a prohibition against the use of AI to prevent or reduce the exchange of illegal content generated by means of AI tools, will eventually become absurd.

Currently, the EU is about to adopt a legal framework that requires platform providers to monitor communication and moderate content to protect children against sexual exploitation and abuse online. This so called 'chat control' has sparked quite a debate, but such a general limitation of privacy can very well be deemed necessary for the protection of the child and, hence, proportionate under the General Data Protection Regulation (GDPR) which in turn operationalises Article 52 of the EU Charter. Perhaps, the very automation of monitoring makes it easier for humans to accept the intrusion of privacy, insofar as it abstracts the supervision

of content sharing from human social control. Furthermore, platform providers that become aware of possible infringements of rights and interests already have an obligation to act under the e-commerce directive and Digital Service Act (DSA). On that note, automation blurs the concepts of 'monitoring' and 'awareness', since it is becoming increasingly unlikely that an AI system would be 'unaware' of content shared by the end-users. Ultimately, the legislator must determine for what purpose data may be processed by the AI system.

As uploaded and downloaded information is linked to a specific account, it typically contains personal data. Hence, it will be subject to the prohibition against automated decisions under Article 22 of the GDPR. According to paragraph 2 of that provision, exemptions can be made if the deployment of AI is specified in legislation at the EU level or national level (since consent and contractual reasons will rarely apply). That is why the EU is about to adopt a legal framework for chat control regarding child sexual abuse. However, the same rationale regarding limitations of privacy and AI-decisions applies to monitoring and content moderation for the protection of all kinds of private rights and social interests. For instance, repression of dis- and misinformation that threatens democracy may also justify limitations of the right to privacy in order to restrict or rather define the freedom of expression online. **Whereas the deployment of AI to shape online customs with ramifications for society at large needs to be justified by legislation, the deontic modalities of legislation need to be transposed into algorithms**. Indeed, the red lines for freedom of expression must be identified by humans at some level since it is challenging already to identify the legal limits of the freedom of expression in order to protect other rights and interests, `moral judgements that are legally acceptable must be left to humans`.

At the outset, the concept of freedom of expression encompasses virtually all kinds of human expression. It implies that humans can receive and impart uncomfortable and even offensive views, act out of kilter, share ugly images and make noise without conveying any intelligible information or 'music'. However, noise and ugly images can constitute molestation and even torture depending on the context, veracious information may affect persons in unacceptable ways and deceit is rarely justifiable. Classical interfaces that limit the freedom of expression are infringement of copyrights in terms of plagiarism, infringement of trade mark rights in terms of damage to commercial reputation, communication that amounts to slander or defamation of natural persons, as well as discrimination and sexual abuse, threats and undue influence, and instigation to many crimes defined in national legislation. However, the classification of expression is blurred by the ambiguity of human communication, and a seemingly unacceptable expression may be covered by artistic or scientific freedom, as well as constituting satire or irony, or any other form of fair messaging such as social criticism. Emerging technology and training will most likely make it possible for AI to calibrate decisions and casuistically identify the red lines of law, within which humans are free to make value judgements. Perhaps different standards should apply for different communities, and hence, online access to a community will determine the scope of the freedom of expression via the online platform.

If accepting AI monitoring of human communication via online platforms and policing of human rights and interests online, the model for enforcement of the rights and interests changes from reactive interventions by human institutions to proactive assessment and moderation of human interaction. Consequently, sanctions may be less categorical and more formative for social interaction, as risk assessments may not have to result in content moderation or conclusive take down. Instead of prohibiting the use of an AI system such as Grok in the European cloud or sphere, AI can detect and casuistically take down unacceptable content generated by the Grok algorithm.

Instead of 'sugar coating' the ability of the AI system to classify communication, transparency could be promoted by a signalling system akin to traffic lights and 'watermarking' of content. Conversely, clear criminal offences are preferably directly reported to law enforcement agencies. After all, the purpose of legislation is to regulate human behaviour and the imperfection of current enforcement models is a problem as opposed to a factor that makes the normative measures proportionate. If AI can be deployed to make the enforcement of legal standards more efficient, the normative impact of law would increase, which in turn reduces the need for excessive sanctions as a deterrent. As long as the deployment of AI is conditioned on legislation, man can shape cultures by means of AI instead of AI shaping human cultures in terms of streamlining and standardising human expression. Diversification and contextualisation are the challenge ahead for the socio-computer scientist.

3 A DEMOCRACY SHIELD IN AGENTIC AI

AiCrit operationalizes the 'social turn' by shifting governance from the content layer to the structural layer. It implements Safety-by-Architecture (SbA) as a foundational alternative to post-hoc safety paradigms. Unlike Safety-by-Design, which integrates considerations throughout the development lifecycle, SbA embeds hard governance mechanisms directly into the computational architecture of AI agents. Drawing from systems thinking (leverage points in complex systems) and hardware security principles (rendering certain vulnerabilities impossible by design), SbA enforces structural invariants on reasoning and behaviour, reducing reliance on fragile filters prone to jailbreaking or bypass. AiCrit realizes SbA through two interlocking pillars that shift governance from content-level filtering to the structural layer. As Mavračić (2025) notes, "an AI agent can only be assured if its governing requirements are considered part of the deployed model itself". By treating these requirements as structural invariants, AiCrit moves toward accountable autonomy at scale.

3.1 Topologically Constrained Reasoning

AiCrit structures agent reasoning as a Directed Acyclic Graph (DAG), where nodes represent discrete, immutable steps. For example, a process that contains proposal generation, legal/policy check, impact assessment, stakeholder review, provenance anchoring, and final output has directed edges with strict dependencies. Topological sorting ensures forward-only execution: a node activates only after all prerequisites complete and return a required boolean validation (e.g. "True" from critical gates). If any gate fails,

the process terminates, escalates, or logs for human review. This Audit Gate-design embeds democratic principles into the architecture, addressing vulnerabilities in monolithic LLM reasoning:

- Traditional single-model agents risk self-deception to complete a task. DAG enforces a "clean separation of responsibilities" where proposing agents are isolated from validation nodes, preventing the system from "re-explaining" a rejected proposal without unobserved internal overrides.
- Hallucinations or misalignments in opaque systems are often untraceable. In AiCrit, discrete dependencies create deterministic audit trails. A post-mortem analysis can pinpoint exactly where failure occurred (e.g. "Ethical Impact Node flagged societal risk"), enabling targeted accountability rather than vague blame on the model.
- Conventional alignment prioritizes the developer–user dyad, but may ignore broader societal harm (e.g. generating persuasive but deceptive political content). AiCrit mandates a citizen/stakeholder proxy node as a required gate for open-mindedness, forcing explicit simulation of impacts on voiceless groups. Societal externalities become a technical dependency, operationalizing the Democracy Shield.
- RLHF-style training relies on subjective "niceness". Topological constraints act like a physical turnstile. Passage requires a valid "ticket" from a Policy Card gate, shifting safety from probabilistic alignment to verifiable process.

In runtime orchestration, gates integrate with the JSON Schema policy registry for rule lookup and Confidential and Traceable Retrieval-Augmented Generation (CT-RAG) for provenance checks. DAGs excel in compliance-critical domains by prioritizing auditability and normative protection. This draws from emerging governed orchestration patterns (e.g. validator-gated execution and policy-aware DAGs in enterprise agentic systems).

3.2 The Citizen Proxy as a Mandatory Gate

To operationalize the Protector of the Voiceless principle, AiCrit introduces a dedicated Citizen Proxy Agent. This is an adversarial and critical thinking agent that evaluates proposals against a JSON-encoded Normative Registry. This is a set of formal policy constraints representing societal interests (e.g. non-deception, or local impact). AiCrit does not treat a "Stop" gate as a terminal failure but as a trigger for reflection. Unlike standard agents that optimize for user-defined efficiency, the Citizen Proxy is architecturally mandated to simulate "normative externalities". Because the DAG is acyclic, if the Citizen Proxy returns a boolean "False" based on a policy violation, the process terminates immediately, preventing the system from "re-justifying" its way around a societal red line.

4 ANCHORING, PROVENANCE, AND VERIFIABLE AUDITABILITY

To ground reasoning in objective sources and enable long-term, tamper-resistant scrutiny, AiCrit combines Confidential and Traceable Retrieval-Augmented Generation with on-chain commitments to Chromia, ChromaWay's relational blockchain platform. CT-RAG maintains cryptographically linked provenance (e.g. hash-chained citations, signed metadata) between generated outputs and retrieved evidence, while conclusions follow formal premise-conclusion structures where feasible, shifting inference toward auditable deduction.

Key artifacts may be selectively committed to Chromia for immutable, queryable verification:

(1) A cryptographic hash of the reasoning path (e.g. Legal Check → Ethical Impact → Stakeholder Proxy) proves adherence to sequential, non-bypassable gates, providing court-admissible evidence, in democratic contexts like public procurement, that no steps were skipped or altered.
(2) Hashes of source documents (via CT-RAG) anchor claims to exact versions (e.g. a specific procurement document), ensuring verifiability as immutable as the law.
(3) Pass/fail outcomes from critical nodes (e.g. stakeholder proxy) form searchable, aggregated logs, enabling public macro-transparency without exposing private details.
(4) The final output links to the input's hash, preventing post-facto tampering. Mismatches immediately reveal irregularities.

Selective hashing minimizes latency and cost overhead, maximizing auditability. Authorized stakeholders can independently verify policy compliance, making governance transparent, accountable, and resistant to manipulation. Combined with topological constraints, this shifts opaque inference to a verifiable, auditable process

While traditional safety paradigms rely on the model's internal values, AiCrit's SbA treats societal alignment as an external infrastructure requirement. By encoding policy into a JSON Schema registry and enforcing it through a DAG-based orchestrator, we transform ethics from a probabilistic output into a deterministic dependency. The 'Democracy Shield' becomes an invariant of the system's execution path.

5 AICRIT VERSUS TRADITIONAL APPROACHES

AiCrit reinterprets several core concepts from normative and interaction-oriented multi-agent systems research within the specific context of agentic generative AI based on large language models, while directly addressing the limitations of current Big Tech-dominated systems that often function as opaque black boxes. In traditional agentic AI:

(1) norms and roles are typically enforced through:
  (a) hardcoded prompts, absent formal representation,
  (b) emergent behavior from fine-tuning, or
  (c) a combination thereof.
(2) coordination depends on ad-hoc natural language exchanges or simple sequential workflow,
(3) trust and auditability remain limited, with outputs largely opaque and explainability restricted to post-hoc techniques, and
(4) theory-of-mind capabilities stay rudimentary, usually limited to basic persona simulation in prompts.

In contrast, AiCrit embeds governance at the architectural level by introducing explicit formal policy encodings through a JSON Schema registry that serves as architectural primitives with clearly defined normative scopes to ensure constraint adherence; enforcing dependency-ordered execution via Directed Acyclic Graphs with

topological constraints to prevent circular justification loops that enable deceptive rationalizations. This counters opacity through:

(1) verifiable grounding using CT-RAG combined with immutable reasoning logs stored on the Chromia relational blockchain, thereby creating tamper-resistant and queryable audit trails, and
(2) extending perspective-taking by incorporating dedicated Citizen and Stakeholder proxies that allow the system to reason explicitly about critical points of view, societal impacts and normative externalities beyond the traditional developer–user dyad.

To illustrate Safety-by-Architecture (SbA), consider an environmental assessment where a traditional AI might subtly favor a vendor due to narrow prompts or biased data. In the AiCrit framework, the Citizen Proxy node acts as a mandatory gate, checking if the input of an indigenous group was included in the assessment. If an audit gate returns a 'Stop' or 'Escalate' status, AiCrit initiates a collaborative revision cycle. Monolithic agents might 're-explain' away the omission to satisfy a reward signal (user tampering), instead AiCrit forces the proposer node to ingest the missing stakeholder perspective as a new technical dependency. This operationalizes the 'topology of open-mindedness', providing the 'intermediary pauses' necessary for the human executor to exercise reflective endorsement before reaching a final decision.

Via a transparent, normatively grounded, and aligned with the needs of indirect stakeholders, AICRIT supports the broader social-contextual turn in multi-agent systems research.

6 CONCLUSION

Our proposal, AICrit, demonstrates that the "social-contextual turn" in multi-agent systems research is technically feasible through Safety-by-Architecture (SbA). By implementing topological constraints (via DAG-enforced Audit Gates), formal policy registries (JSON Schema-encoded normative boundaries), and dedicated Citizen Proxies as mandatory reasoning nodes, we shift the burden of alignment from fragile, individual-model training (e.g. easily bypassed RLHF) to verifiable structural invariants in the system architecture itself.

This approach honours the mandate of international law, and the EU Charter, DSA, and AI Act to protect the "voiceless". They are the indirect stakeholders impacted by AI decisions but excluded from the developer-user dyad. Ultimately, the MAS community faces a choice between perpetuating a regime of megatech tutelage, where opaque, efficiency-driven silos amplify biases and deception, or embrace more multi-disciplinary thinking to uncover and empowering the hidden and the silent. By targeting high-leverage points in agent design (business models, normative externalities, verifiable provenance etc.), we can transform agentic AI from a tool of potential surveillance or exclusion into a genuine Democracy Shield.

The future of autonomous agents must not reflect monopoly power, but rather our shared societal values and fundamental rights. Then "Code is Law" becomes an affirmation of democratic governance rather than an observation of its erosion.