Who Guards the Agents? Bias, Fairness, and Oversight in Agentic AI Systems

Francisco-Javier Rodrigo-Ginés
Universidad Nacional de Educación a Distancia
Madrid, Spain
fjrodrigo@invi.uned.es

Flor Miriam Plaza-del-Arco
Leiden Institute of Advanced Computer Science
Leiden University, Netherlands
f.m.plaza.del.arco@liacs.leidenuniv.nl

***Abstract*—Agentic AI systems (autonomous agents powered by large language models that reason, use tools, and collaborate with other agents) are rapidly moving from research prototypes to deployed applications. Unlike static AI models, agentic systems make sequential decisions, select information sources, and take consequential actions with limited human oversight. This autonomy introduces trustworthiness challenges that existing frameworks, designed for single-model systems with human-in-the-loop supervision, do not adequately address. Biases encoded in the underlying language models are not merely reproduced but *amplified* through multi-step reasoning chains, autonomous tool selection, and multi-agent coordination. This paper examines the specific trustworthiness challenges that agentic AI introduces and proposes a comprehensive framework to address them. Our contributions include: (1) a systematic analysis of how bias propagates and amplifies through autonomous agent reasoning, drawing on research in media bias detection and emotion analysis; (2) a formal probabilistic model of bias propagation in multi-step agent chains; (3) a seven-layer architecture for agentic systems that integrates ethics and governance layers by design; (4) a lifecycle bias detection pipeline spanning pre-deployment, runtime, and post-deployment phases; (5) design principles for emotion-aware agents that avoid perpetuating demographic stereotypes; and (6) a compliance analysis of four major agentic frameworks against our trustworthiness requirements. We illustrate the framework through three application scenarios and compare our approach with existing trustworthy AI instruments.**

***Index Terms*—Agentic AI, Trustworthy AI, Bias Detection, Fairness, Large Language Models, Emotion Analysis, Human Oversight, Accountability**

I. Introduction

Artificial intelligence is undergoing a fundamental architectural shift. Where earlier systems processed inputs and returned outputs within a single inference step, a new generation of *agentic* systems autonomously plans multi-step strategies, invokes external tools, retrieves information from diverse sources, and coordinates with other agents to accomplish complex objectives [1]–[3]. Frameworks such as AutoGen [4] and MetaGPT [5] have made it possible to deploy collaborative agent architectures in domains ranging from software engineering and scientific research to healthcare triage and financial analysis. More recently, platforms such as LangChain [6] and CrewAI [7] have further lowered the barrier to building and deploying multi-agent systems, accelerating adoption across industry sectors. Recent surveys identify hundreds of agentic systems in active development, most built on top of large language models (LLMs) whose capabilities serve as the cognitive backbone of autonomous decision-making [8], [9].

The scale and speed of this deployment are difficult to overstate. In less than two years, the field has moved from single-agent prototypes that could chain a handful of tool calls to sophisticated multi-agent orchestrations capable of conducting research, writing code, managing workflows, and making decisions that affect real people. Every major technology company now offers an agentic AI platform: Anthropic's Claude, OpenAI's Assistants API, and Google's Gemini agents are already integrated into enterprise workflows where they autonomously handle customer service, document processing, and strategic analysis. The adoption spans virtually every sector of the economy. In healthcare, agentic systems triage patient inquiries, synthesize medical literature, and draft clinical summaries. In finance, they monitor regulatory filings, generate risk assessments, and execute trading strategies. In legal services, they review contracts, identify compliance gaps, and draft legal memoranda. In education, they serve as personalized tutors that adapt to individual learning trajectories. In hiring, they screen resumes, evaluate candidate responses, and rank applicants. Each of these domains involves consequential decisions about human welfare, and in each, agentic systems are being deployed with varying degrees of human oversight.

The breadth of adoption is matched by the depth of investment. Major technology firms have committed billions of dollars to agentic AI infrastructure, and the venture capital ecosystem has channeled substantial funding into startups building agent-based tools for specific industries. Open-source frameworks have proliferated, with LangChain, CrewAI, AutoGen, and MetaGPT collectively accumulating tens of thousands of contributors and millions of downloads. Enterprise adoption surveys consistently report that a majority of organizations are either deploying or actively piloting agentic AI systems in production environments. This is not a technology confined to research laboratories; it is a technology being embedded into the operational infrastructure of organizations that serve millions of people daily. The implications for trustworthiness are proportional to this scale.

This transition from tool to agent carries profound implications for trustworthiness. LLMs are known to encode systematic biases, including gender stereotypes in emotion

attribution [10], religious stigmatization [11], and political framing in information synthesis [12]. Alignment techniques such as RLHF [13] and constitutional AI [14] reduce but do not eliminate these biases. In a static assistant, a biased output is a single event that a human user can evaluate and discard. In an agentic system, bias compounds. An agent that autonomously selects information sources, reasons over retrieved documents, and generates a synthesis can amplify subtle training-data biases at each step of its reasoning chain. When multiple agents collaborate, the amplification multiplies further, as each agent's biased intermediate outputs become another agent's inputs.

The compounding dynamics deserve careful attention because they operate across multiple, interacting channels simultaneously. Consider the domain of hiring. A recruitment agent tasked with screening candidates might autonomously query a resume database, select evaluation criteria, score applicants, and generate a shortlist. At the source selection step, the agent might favour databases that over-represent certain demographic profiles in their indexed populations. At the criteria selection step, it might weight qualifications that correlate with socioeconomic status rather than job performance, reflecting patterns in its training data. At the scoring step, it might interpret identical qualifications differently depending on contextual cues (names, institutions, geographic locations) that serve as demographic proxies. At the shortlisting step, it might impose implicit diversity thresholds that are either too lax or too aggressive, depending on what patterns it has learned from historical hiring outcomes. No single step is transparently discriminatory, yet the cumulative effect can systematically disadvantage entire demographic groups. In a multi-agent variant (where separate agents handle sourcing, screening, interviewing, and ranking) each handoff introduces an additional opportunity for bias to enter and accumulate, while the distributed architecture makes it harder for any single auditor to trace the full decision chain. Similar compounding dynamics arise in healthcare (where an agent that retrieves medical literature, interprets symptoms, and recommends treatment plans may propagate biases in clinical research that historically under-represent minority populations), in criminal justice (where risk assessment agents may inherit and amplify the racial disparities embedded in historical crime data), and in content moderation (where agents that classify, escalate, and act on flagged content may disproportionately suppress speech from marginalized communities) [15], [16].

A critical dimension of the current moment is the widening gap between the pace of agentic AI deployment and the development of governance mechanisms adequate to the technology's capabilities. Regulatory frameworks, by their nature, evolve slowly: the EU AI Act [17], years in development, was designed primarily with static predictive models and classification systems in mind. Its risk-based categorization assigns high-risk status based on application domain (hiring, credit scoring, law enforcement) rather than on the degree of autonomous reasoning an AI system performs, meaning that a simple logistic regression classifier used in hiring is subject to more stringent oversight obligations than a sophisticated multi-agent system that autonomously plans and executes research tasks, provided the latter falls outside the enumerated high-risk categories. Voluntary governance instruments, including the NIST AI Risk Management Framework [18] and ISO/IEC 42001 [19], offer process-oriented guidance but lack the specificity required to address the emergent behaviours, multi-step reasoning chains, and distributed accountability structures that characterize agentic systems. Meanwhile, industry self-regulation has focused primarily on alignment of base models (through RLHF, constitutional AI, and red-teaming) rather than on the systemic risks that arise when aligned models are deployed as autonomous agents. The result is a governance vacuum: agentic AI systems are being deployed into high-stakes domains at a pace that outstrips the capacity of existing frameworks to ensure their trustworthiness. Bridging this gap requires not merely extending existing frameworks but developing new conceptual and architectural approaches that are native to the agentic paradigm.

Existing trustworthy AI frameworks were not designed for this paradigm. The EU's Ethics Guidelines for Trustworthy AI [20] assume meaningful human oversight at decision points. The EU AI Act [17] classifies risk based on application domain, not on the degree of autonomous reasoning an AI system performs. The NIST AI Risk Management Framework [18] provides governance processes but does not address the emergent behaviors that arise when agents interact. These frameworks offer essential principles, but they leave critical gaps when applied to systems that act, reason, and collaborate with minimal human intervention.

This paper addresses three research questions that arise from these gaps:

1) **Bias propagation in autonomous reasoning.** How can bias be detected and mitigated in AI systems that reason, act, and collaborate autonomously across multi-step decision chains? Existing bias detection methods are designed for single-model, single-inference settings: they evaluate a model's outputs against a fixed set of inputs and measure demographic disparities in predictions [15], [21]. These methods do not capture the compounding dynamics that arise when an agent makes a sequence of interdependent decisions, each conditioned on the (potentially biased) outputs of previous steps. Nor do they account for the bias introduced through autonomous tool selection, where the agent's choice of information sources may itself be a vector of systematic distortion [12]. Addressing this question requires both a formal understanding of how bias propagates through agent reasoning chains and practical detection mechanisms that can operate at each stage of the chain, including the inter-agent handoff points where bias amplification is most acute.
2) **Transparency and accountability in distributed agent architectures.** What architectural principles ensure transparency and accountability when decision processes span multiple agents and external tools? In a single-

model system, transparency can be achieved through model cards [22], datasheets [23], and post-hoc explanation methods [24]. In a multi-agent system, the decision chain traverses multiple models, multiple tool invocations, and potentially multiple organizational boundaries, making it unclear where transparency obligations begin and end, who is accountable for emergent system-level behaviours, and how an affected individual can contest a decision whose causal chain spans a distributed architecture. These challenges are not merely technical; they are fundamentally architectural, requiring that transparency and accountability be designed into the system's structure rather than retrofitted onto its outputs.

3) **Emotion-aware agents and demographic stereotyping.** How can emotion-aware agents be designed to recognize and respond to human emotions without perpetuating demographic stereotypes? Emotion recognition is increasingly integrated into agentic applications in mental health support [25], customer service, and education. Yet LLMs systematically reflect gendered [10] and religious [11] stereotypes in how they attribute emotions, and the scientific basis of categorical emotion recognition itself remains contested [26]. When an agent acts on stereotyped emotional inferences (escalating a support case, adjusting communication tone, or modifying recommendations) the bias moves from perception to action, with potentially discriminatory consequences. Designing emotion-aware agents that are both empathetic and fair requires integrating insights from emotion analysis research [27] with bias-aware architectural safeguards.

We make the following contributions:

1) A systematic analysis of trustworthiness challenges *specific* to agentic AI, demonstrating that autonomous reasoning, tool selection, and multi-agent coordination create qualitatively new bias risks that static-model frameworks do not capture. This analysis draws on established research in media bias detection [12] and emotion attribution [10] to ground the agentic trustworthiness problem in empirically documented bias patterns.
2) A formal probabilistic model of bias propagation that quantifies how bias compounds through multi-step reasoning chains and multi-agent handoffs, providing a theoretical foundation for understanding the amplification dynamics. The model distinguishes between bias inheritance, amplification, and introduction at each processing step, and derives closed-form expressions for expected bias levels in both single-agent and multi-agent configurations.
3) A seven-layer architecture for agentic systems that integrates dedicated ethics and governance layers, ensuring that bias checking and value alignment are structural components rather than afterthoughts. The architecture specifies where in the agent's processing pipeline bias checks should occur, how inter-agent communications should be monitored, and how human oversight should be integrated through a tiered autonomy model.
4) A lifecycle bias detection pipeline that spans pre-deployment assessment, runtime monitoring, and post-deployment auditing, incorporating techniques from media bias detection [12] and emotion attribution analysis [10]. The pipeline adapts established bias detection methodologies (including lexical bias analysis, framing detection, and omission analysis) to the specific challenges of agentic reasoning chains.
5) Design principles for emotion-aware agents grounded in empirical research on gendered and religious stereotyping in LLMs [10], [11], providing actionable guidelines for building agents that respond to human emotions without reinforcing harmful stereotypes. These principles address the full pipeline from emotion perception to emotion-informed action, with specific safeguards against stereotyped attribution.
6) A compliance analysis evaluating four major agentic AI frameworks (AutoGen, LangChain, CrewAI, MetaGPT) against our trustworthiness requirements, identifying systematic gaps in current implementations and providing concrete recommendations for framework developers and deployers.

The remainder of the paper is organized as follows. Section II reviews the landscape of agentic AI, trustworthy AI principles, and bias in LLMs. Section III analyzes the specific trustworthiness challenges that agentic systems introduce and presents a formal model of bias propagation. Section IV presents our framework, including the layered architecture, bias detection pipeline, emotion-aware design principles, and tiered autonomy model. Section V evaluates existing agentic frameworks against our trustworthiness requirements and presents an empirical validation protocol. Section VI applies the framework to three illustrative scenarios. Section VII compares our approach with existing frameworks, acknowledges limitations, and identifies open challenges. Section VIII concludes.

II. Background and Related Work

A. From Assistants to Autonomous Agents

The term "agentic AI" refers to systems that go beyond single-turn question answering to exhibit four properties that distinguish them from traditional LLM assistants: *autonomy* (operating with minimal human intervention), *reactivity* (perceiving and responding to environmental changes), *proactivity* (pursuing goals through multi-step planning), and *social ability* (collaborating with other agents and humans) [8], [9]. These properties are not merely incremental improvements; they represent a qualitative shift in how AI systems interact with the world, transforming language models from passive oracles that respond to isolated queries into active participants that shape their information environments.

The technical foundations of this shift emerged rapidly. Yao et al. [1] demonstrated that interleaving reasoning traces with

action steps (the ReAct paradigm) enables LLMs to solve complex tasks that require both deliberation and environmental interaction. This paradigm is significant because it couples the model's internal reasoning with external state changes: the agent does not merely think about what to do but executes actions whose outcomes feed back into subsequent reasoning steps, creating a closed loop between cognition and environment. Schick et al. [2] showed that language models can learn to invoke external tools autonomously, extending their capabilities beyond text generation to include calculation, search, database queries, and API calls. Tool use is a pivotal capability because it allows agents to exceed the knowledge boundaries of their training data and to produce effects in external systems (sending emails, modifying files, executing transactions) that have real-world consequences beyond the generation of text. Park et al. [3] created generative agents that simulate believable human behavior over extended time horizons, maintaining persistent memory and social relationships. Their work demonstrated that LLM-based agents can maintain coherent identities, form plans based on accumulated experience, and engage in emergent social dynamics, including cooperation, information sharing, and coordination, without explicit programming of these behaviors. Multi-agent frameworks such as AutoGen [4] and MetaGPT [5] formalized collaborative architectures in which specialized agents divide labor, critique each other's outputs, and converge on solutions through structured dialogue. MetaGPT, in particular, demonstrated that assigning agents role-specific standard operating procedures (mimicking the division of labor in human software engineering teams) significantly improves the quality and coherence of collaborative outputs.

The technical capabilities of modern agentic systems now span a wide range of modalities and interaction types. Tool use has expanded from simple calculator and search invocations to encompass code execution (where agents write, test, and debug software), web browsing (where agents navigate websites, fill forms, and extract information from dynamic pages), file system manipulation (where agents create, read, and modify documents), and API orchestration (where agents compose calls to multiple external services to accomplish complex workflows). Code execution is particularly consequential because it gives agents the ability to perform arbitrary computation, transforming them from text processors into general-purpose problem solvers that can analyze data, generate visualizations, train machine learning models, and automate administrative tasks. Web browsing capabilities introduce additional trustworthiness challenges because they expose agents to the full complexity and adversarial dynamics of the open internet, where misinformation, manipulation, and bias in online content can directly influence agent reasoning.

Deployment contexts have expanded with equal speed. In enterprise settings, agentic systems are being integrated into customer service workflows (where they handle multi-turn support conversations, access internal knowledge bases, and escalate to human agents), document processing pipelines (where they extract, summarize, and cross-reference information across large document collections), strategic analysis (where they gather market intelligence, synthesize competitive landscapes, and generate reports), and software development (where they write, review, and deploy code with minimal human supervision). In consumer-facing applications, agents serve as personal assistants that manage calendars, draft communications, conduct research, and make purchasing decisions on behalf of users. In research contexts, agentic systems are being deployed to conduct literature reviews, generate hypotheses, design experiments, and write scientific manuscripts. Each of these deployment contexts carries distinct trustworthiness implications: enterprise deployments raise questions about liability and the unauthorized disclosure of proprietary information; consumer deployments raise questions about manipulation, consent, and the asymmetry of information between agent and user; research deployments raise questions about scientific integrity, the hallucination of citations, and the distortion of evidence synthesis.

The speed of adoption has been remarkable by any measure. In less than two years, the field has moved from single-agent prototypes that could chain a handful of tool calls to sophisticated multi-agent orchestrations deployed in production environments that serve millions of users. Major technology companies have released agent-building platforms (Anthropic's Claude with tool use, OpenAI's Assistants API, Google's Gemini agents) that are already integrated into enterprise workflows. Open-source frameworks have proliferated in parallel. LangChain [6] provides a modular framework for building agent applications with retrieval-augmented generation, tool use, and chain-of-thought reasoning, and has become a de facto standard for rapid prototyping. CrewAI [7] introduces role-based multi-agent collaboration with configurable autonomy levels, enabling developers to assemble teams of specialized agents with minimal code. These frameworks have dramatically lowered the barrier to deploying agentic systems in production environments: what previously required months of custom engineering can now be accomplished in days, often by developers with limited understanding of the underlying model's biases and limitations. This democratization of agentic capabilities is, in itself, a governance challenge: the same accessibility that enables innovation also means that consequential agentic systems are being deployed by organizations without the expertise or infrastructure to evaluate their trustworthiness.

The ecosystem is growing but fragmented. No dominant standard governs how agents communicate with one another, how tools expose their capabilities and constraints, or how multi-agent systems report their decision processes to human overseers. Rodrigo-Ginés [28] identifies the absence of standardized communication protocols and interoperability layers as a critical barrier, drawing an analogy to the pre-OSI era of computer networking, when incompatible proprietary protocols made it impossible to build reliable, auditable, cross-platform communication systems. This fragmentation complicates governance in multiple ways. Without shared architectural abstractions, it is difficult to specify where bias

checks should occur, what transparency obligations apply, or how accountability should be distributed across a multi-agent system. Each framework implements its own conventions for agent-to-agent messaging, tool registration, memory management, and error handling, meaning that governance mechanisms developed for one framework do not transfer to another. The absence of standardized audit interfaces makes it nearly impossible for regulators or external auditors to inspect multi-agent systems that span multiple frameworks and organizational boundaries. Just as network security was impossible to enforce systematically before the standardization of networking protocols, agentic AI governance will remain ad hoc and framework-specific until common architectural abstractions emerge.

B. Trustworthy AI: Principles and Regulation

The concept of trustworthy AI has been codified through several influential frameworks, each approaching the challenge from a distinct normative, regulatory, or operational perspective. Understanding both their contributions and their limitations with respect to agentic systems is essential for motivating the framework proposed in this paper.

The EU's High-Level Expert Group on AI articulated seven requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal well-being, and accountability [20]. These requirements have shaped subsequent policy instruments and have been echoed, with variations, in over 80 national and international AI ethics guidelines surveyed by Jobin et al. [29]. The EU Ethics Guidelines are foundational in establishing that trustworthiness is not a single property but a multidimensional requirement encompassing technical, ethical, and social dimensions. However, the guidelines' operationalization of human oversight assumes a supervision model in which humans occupy meaningful decision points in the AI pipeline, reviewing outputs, approving actions, and intervening when necessary. Agentic systems disrupt this assumption fundamentally: an agent that autonomously decomposes a complex task into dozens of subtasks, executes each through tool invocations, and synthesizes the results does not naturally present the kind of discrete decision points at which human oversight is assumed to occur. The guidelines acknowledge the importance of human agency but provide no mechanism for maintaining it when the AI system itself determines the sequence, timing, and granularity of its actions.

On the regulatory front, the EU AI Act [17] establishes a risk-based classification framework in which high-risk AI systems face conformity assessments, documentation requirements, and post-market surveillance. The Act represents the most comprehensive legislative effort to regulate AI to date and introduces several mechanisms relevant to trustworthiness, including mandatory risk management systems, data governance requirements, transparency obligations toward users, and human oversight provisions. However, the Act's risk classification is organized around application domains (biometric identification, critical infrastructure, employment, law enforcement) rather than around the degree of autonomous reasoning an AI system performs. This domain-based classification creates a regulatory gap for agentic systems: a highly autonomous agent that operates in a domain not classified as high-risk (for example, an agent that autonomously conducts market research or manages social media content) may escape the Act's most stringent requirements despite exhibiting the kind of autonomous, multi-step reasoning that amplifies bias in the ways documented in Section III. Furthermore, the Act's conformity assessment procedures are designed for systems that can be evaluated as bounded artifacts before deployment. Agentic systems that learn from interactions, adapt their strategies over time, and exhibit emergent behaviors when combined with other agents are not well captured by a pre-deployment conformity assessment paradigm, because their risk profile changes dynamically in ways that cannot be fully anticipated at the time of assessment. The Act's provisions for post-market surveillance partially address this limitation, but the specific monitoring mechanisms needed for agentic systems (such as tracking bias amplification across multi-step reasoning chains or auditing inter-agent handoffs) are not specified.

The NIST AI Risk Management Framework [18] provides a voluntary governance structure organized around four core functions: govern (establishing policies and processes), map (identifying and categorizing AI risks in context), measure (quantifying risks using appropriate metrics), and manage (prioritizing and acting on identified risks). The NIST framework is notable for its process orientation: rather than prescribing specific technical requirements, it provides a structured methodology for organizations to develop their own risk management practices. This flexibility is a strength in that it can accommodate diverse AI systems and organizational contexts, but it is also a limitation for agentic systems. The framework's guidance on risk identification and measurement does not address the specific dynamics of autonomous multi-agent systems, such as emergent behaviors arising from agent interactions, bias amplification through sequential reasoning, or the accountability challenges created by distributed decision-making across multiple agents. The "map" function, which calls for contextualizing AI risks, does not distinguish between systems that make single inferences and systems that execute extended autonomous reasoning chains, despite the qualitatively different risk profiles these architectures present. The "measure" function provides general guidance on selecting and applying metrics but does not specify how fairness metrics should be adapted for systems whose outputs depend on dynamic, multi-step decision processes rather than single input-output mappings.

ISO/IEC 42001 [19] introduces a management system standard for organizations developing or using AI, providing a certification pathway for AI governance. Modeled on established management system standards (such as ISO 27001 for information security), it requires organizations to establish an AI management system that includes policy, planning, support, operation, performance evaluation, and improvement

processes. ISO 42001 is valuable because it provides an auditable governance structure that can be integrated with existing organizational management systems. Its limitation for agentic systems is that it focuses on organizational governance processes rather than on the technical architecture of the AI systems themselves. An organization can achieve ISO 42001 certification for its AI governance processes while deploying agentic systems that lack the architectural mechanisms (such as the Ethics and Governance layers proposed in this paper) needed to operationalize those governance processes at the system level. The standard does not prescribe how AI systems should be architected to support trustworthiness, leaving a gap between organizational governance commitments and their technical implementation in autonomous agent architectures.

Floridi et al. [30] proposed a complementary ethical framework organized around five principles: beneficence (AI should promote well-being), non-maleficence (AI should not cause harm), autonomy (AI should preserve human self-determination), justice (AI should promote fairness and prevent discrimination), and explicability (AI should be transparent and accountable). The AI4People framework bridges bioethical tradition and AI governance, providing a principled vocabulary for evaluating AI systems against established ethical norms. Its contribution lies in offering a coherent normative foundation that transcends specific regulatory contexts. However, the framework operates at a high level of abstraction: it identifies what ethical principles AI systems should satisfy but does not specify how those principles should be operationalized in concrete architectures or governance mechanisms. The principle of autonomy, for instance, calls for preserving human self-determination in the face of AI systems, but does not address the specific challenge of maintaining meaningful human autonomy when interacting with an agent that makes dozens of intermediate decisions autonomously, each shaping the information environment in which the human ultimately acts. The principle of justice calls for fairness, but does not address how fairness should be measured, maintained, or enforced in systems where bias compounds through multi-step reasoning chains rather than manifesting in single outputs.

These frameworks provide essential normative foundations. However, they share a common assumption: the AI system under governance is a relatively static artifact (a model that receives inputs and produces outputs, with humans positioned at meaningful decision points). Agentic AI disrupts this assumption in three ways. First, agents act autonomously over extended sequences of decisions, reducing opportunities for human checkpoint intervention and making the "human oversight" requirement aspirational rather than operational unless new architectural mechanisms are introduced. Second, multi-agent systems exhibit emergent behaviors that are not predictable from the properties of individual agents, making pre-deployment risk assessment insufficient, because the system's risk profile is a function not only of its components but of their interactions, which may vary with deployment context and evolve over time. Third, agents that use external tools and data sources create attribution chains that span organizational boundaries, complicating the assignment of accountability in ways that neither liability law nor existing governance frameworks are equipped to handle. The trustworthy AI principles remain valid, but their operationalization requires rethinking for the agentic paradigm.

C. Bias in Large Language Models

LLMs encode the statistical patterns of their training data, including systematic biases that reflect and sometimes amplify societal inequalities [15], [31]. The relationship between training data and model bias is not a simple one of reproduction; it involves complex dynamics of selection, amplification, and transformation. Training corpora are assembled from internet text, books, and other sources that over-represent certain demographics, geographies, and perspectives while under-representing others. The statistical learning process extracts patterns from these corpora, and because biased patterns are often more consistent and salient than counter-stereotypical ones, models learn to reproduce and even exaggerate them. Furthermore, the curation decisions involved in assembling training data (which sources to include, how to weight them, what filtering to apply) introduce additional biases that are difficult to audit because the composition of proprietary training datasets is rarely disclosed in full.

Gallegos et al. [21] provide a comprehensive survey of bias in LLMs, documenting disparities across gender, race, religion, age, and disability dimensions. Their survey identifies several mechanisms through which bias manifests. *Representational bias* occurs when models associate certain groups with stereotyped attributes (for example, associating women with domestic roles or associating certain ethnicities with criminality). *Allocational bias* occurs when model outputs lead to unequal distribution of resources or opportunities across groups (for example, when an LLM-based resume screener systematically ranks candidates from certain demographics lower). *Quality-of-service bias* occurs when models perform measurably better for some user groups than others (for example, generating more fluent and accurate text in response to standard American English than to African American Vernacular English). Each of these bias types has distinct implications for agentic systems: representational bias affects how agents characterize users and topics; allocational bias affects how agents distribute attention, resources, and actions; and quality-of-service bias means that agents may serve some user populations less effectively than others, even when no overtly discriminatory content is generated.

Weidinger et al. [32] offer a taxonomy of risks from language models that extends beyond bias to encompass discrimination, exclusion, toxicity, misinformation, privacy violations, and environmental harms. Their taxonomy is particularly valuable for the agentic context because it highlights that bias is only one dimension of a broader risk landscape. Agentic systems inherit all of these risks and add new ones: an agent that autonomously generates and publishes content can spread misinformation at scale; an agent that accesses external databases can violate privacy through unauthorized

inference; an agent that executes code can cause environmental or infrastructural damage. Understanding bias in agentic systems therefore requires situating it within this broader risk taxonomy, recognizing that bias interacts with other risk categories in compounding ways.

Two lines of research are particularly relevant to the agentic setting. The first concerns *emotion attribution bias*. Plaza-del-Arco et al. [10] demonstrated that LLMs systematically associate anger with male subjects and sadness with female subjects, reflecting entrenched gender stereotypes. Their methodology, which involved presenting LLMs with emotion-eliciting scenarios where the subject's gender was systematically varied while all other variables were held constant, revealed that the bias is not a marginal effect but a robust and consistent pattern across multiple model families and sizes. In a companion study, Plaza-del-Arco et al. [11] documented religious bias in LLMs, finding that stereotyping and stigmatization vary significantly across faiths, with minority religions receiving disproportionately negative emotional associations. Notably, the pattern of religious bias was not uniform: some faiths were associated with predominantly positive emotions while others were associated with fear, anger, or disgust, reflecting the differential representation of religious groups in the training data. Plaza-del-Arco et al. [27] further mapped the landscape of emotion analysis in NLP, identifying trends, gaps, and future directions that bear directly on the design of emotion-aware agents. Their survey revealed that emotion analysis research has concentrated on a narrow set of basic emotions (primarily Ekman's six categories) and a limited set of languages (predominantly English), leaving significant gaps in the field's ability to model the nuanced, culturally situated emotional expressions that agents encounter in real-world multilingual deployments. These findings are consequential for agentic systems that incorporate emotion recognition: an agent that perceives user emotions through a biased model may take discriminatory actions based on stereotyped attributions, and the narrow emotional vocabulary of current models limits agents' ability to respond appropriately to the full range of human emotional expression.

The second concerns *media bias and persuasion*. Rodrigo-Ginés and Carrillo-de-Albornoz [12] conducted a systematic review of 162 studies on media bias detection, identifying four principal dimensions: lexical bias (loaded word choice that frames subjects and events in evaluative terms), framing bias (selective contextualization that emphasizes certain aspects of a story while downplaying others), omission bias (exclusion of relevant information that would complicate or challenge a preferred narrative), and persuasive techniques (rhetorical strategies such as appeal to authority, appeal to emotion, and loaded language that manipulate reader interpretation). These four dimensions operate at different linguistic levels and interact in complex ways: a text may employ neutral vocabulary (low lexical bias) while systematically framing an issue through a particular ideological lens (high framing bias) and omitting perspectives that would challenge that frame (high omission bias). Rodrigo-Ginés et al. [33] demonstrated that persuasion detection techniques can serve as proxies for media bias identification, revealing a deep structural relationship between the rhetorical strategies used to persuade and the linguistic patterns that characterize biased reporting. Further work on hierarchical modeling [34] showed that propaganda detection shares representational structures with media bias, suggesting that these phenomena exist on a continuum of manipulative communication rather than as discrete categories. For agentic systems that retrieve, synthesize, and present information, these biases compound at every stage: an agent may select biased sources (source selection bias), extract information that reflects the framing of those sources (extraction bias), synthesize the extracted information using the framing patterns absorbed from its training data (generation bias), and present the result as objective (presentation bias), with each step invisible to the end user. The multi-dimensional nature of media bias means that simple keyword-based or sentiment-based bias detection is insufficient; detecting the subtle framing and omission biases that agentic information pipelines can propagate requires the kind of multi-dimensional analysis frameworks developed in the media bias detection literature.

Alignment techniques aim to reduce these biases. Reinforcement learning from human feedback (RLHF) [13] trains models to produce outputs that human evaluators prefer, steering generation away from harmful, biased, or offensive content toward responses that are helpful, harmless, and honest. Constitutional AI [14] uses self-critique against explicit principles to reduce harmful outputs, training models to evaluate and revise their own generations according to a set of constitutional rules. These approaches demonstrably reduce the most egregious forms of bias (overt stereotyping, explicit toxicity, blatant discrimination) but do not eliminate subtler forms of stereotyping, framing, and omission [21]. The subtler biases that survive alignment are, in some ways, more dangerous than the overt biases that alignment removes, because they are harder for users to detect and contest. A model that has been aligned to avoid explicit gender stereotypes may still systematically frame women's achievements in terms of their personal qualities ("dedicated", "nurturing") while framing men's achievements in terms of their professional competence ("innovative", "strategic"), a pattern that is difficult to detect at the level of individual outputs but becomes statistically significant across many interactions. Crucially, alignment is applied to the base model; it does not address the bias amplification that occurs when an aligned model is deployed as an autonomous agent making sequential decisions. Each reasoning step in an agentic chain provides an opportunity for the subtle biases that survive alignment to influence downstream processing, and the compounding dynamics formalized in Section III-B mean that individually small biases can accumulate into significant systematic distortions over the length of a typical agent reasoning chain.

D. Human-Centered AI Design

Human-centered AI (HCAI) positions human needs, values, and capabilities at the center of AI system design [35]. The HCAI paradigm holds that AI systems should augment rather than replace human capabilities, that users should maintain meaningful control over AI behavior, and that system design should be grounded in an understanding of human cognitive and social needs. Amershi et al. [36] distilled 18 guidelines for human-AI interaction, emphasizing the importance of setting appropriate expectations, supporting user control, enabling graceful failure, and making clear what the system can and cannot do. These guidelines were developed for interactive AI assistants; their application to agentic systems requires substantial adaptation, because the defining feature of agentic AI (autonomy) exists in fundamental tension with the defining principle of HCAI (human control).

This tension manifests in several dimensions. The first is the *information asymmetry* problem. Agentic systems that autonomously retrieve, filter, and synthesize information shape the informational landscape within which users make decisions. A user who asks an agent to "research the pros and cons of a particular investment" receives a synthesis that reflects the agent's source selection, framing, and editorial judgment, but the user typically has no visibility into how those choices were made or what information was excluded. The agent's autonomy in information processing undermines the user's ability to exercise independent judgment, because the user is evaluating not the raw information but the agent's curated representation of it. This dynamic is qualitatively different from using a search engine, where the user at least sees a list of sources and can choose which to consult; an agentic system that presents a polished synthesis may obscure the diversity of available perspectives. Meaningful human control in this context requires not merely the ability to accept or reject the agent's output but the ability to understand and interrogate the process by which that output was produced.

The second dimension of tension is *temporal control*. Traditional human-AI interaction occurs in discrete exchanges: the user issues a command, the system responds, and the user evaluates the response before issuing the next command. Agentic systems compress this cycle by executing extended sequences of actions autonomously. An agent that is asked to "organize my inbox and schedule follow-ups" may read dozens of emails, categorize them, draft responses, and schedule calendar events before presenting a summary of actions taken. The user has nominally consented to this activity by issuing the initial instruction, but the granularity of control is coarse: the user did not approve each individual action, and reversing unwanted actions (such as an inappropriately scheduled meeting or an inadvisable reply) may be difficult or impossible after the fact. The HCAI principle of user control requires that users be able to intervene at meaningful points in the agent's decision process, but determining what constitutes a "meaningful point" in a fluid, multi-step agent workflow is itself a design challenge that existing guidelines do not resolve.

The third dimension is the *trust calibration* problem. Amershi et al. [36] emphasize that AI systems should help users calibrate their trust by making system capabilities and limitations transparent. For agentic systems, trust calibration is complicated by the fact that the system's capabilities are not fixed but depend on the tools available, the quality of retrieved information, and the interactions between agents in a multi-agent system. A user may develop appropriate trust calibration for an agent performing routine tasks and then be surprised when the same agent fails dramatically on a task that happens to require capabilities or information outside its reliable operating range. The difficulty of maintaining accurate mental models of agentic systems (what they can do, when they are likely to fail, and how they make decisions) poses a persistent challenge to the HCAI ideal of informed user control.

The tension between autonomy and control is particularly acute in emotion-aware applications. Mohammad [37] developed an ethics sheet for automatic emotion recognition systems, highlighting risks including consent violations (users may not know their emotions are being inferred), stereotyping (systems may attribute emotions based on demographic assumptions rather than actual emotional states), and manipulation (systems may exploit detected emotions to influence user behavior). The consent issue is especially salient in agentic systems: when an agent autonomously adapts its communication style, escalation decisions, or recommendations based on inferred emotions, the user may not be aware that emotion recognition is occurring, much less have consented to it. This creates a form of covert behavioral influence that is difficult to reconcile with the HCAI principle that users should understand and control how AI systems affect their experience.

Barrett et al. [26] challenged the scientific basis of emotion recognition itself, arguing that facial expressions and textual cues do not map reliably onto discrete emotional categories across cultures. Their work, grounded in decades of psychological research, demonstrates that emotion is not a set of natural kinds that can be detected from surface cues but a constructed category whose expression and interpretation vary substantially across individuals, contexts, and cultures. This scientific challenge creates a design dilemma for emotion-aware agents: the technology promises personalized, empathetic interaction, but the underlying science suggests that the mapping from observable cues to internal emotional states is too unreliable to support the kind of confident emotional inference that autonomous action requires. For agentic systems that adapt their behavior to perceived user emotions, these concerns are amplified: an autonomous agent that misattributes emotions based on demographic stereotypes may take consequential actions (escalating a support ticket, adjusting communication tone, modifying recommendations, or altering the information presented) on the basis of flawed emotional inferences. The combination of unreliable emotion detection and autonomous action creates a risk profile that passive emotion recognition systems do not share.

Prior work on hate speech detection has explored human-

centered approaches to bias-sensitive content analysis that illuminate the challenges of deploying such techniques in agentic contexts. Plaza-del-Arco and Nozza [38] demonstrated that zero-shot learning with LLMs can detect hate speech without task-specific training data, but also showed that performance varies across demographic targets, raising fairness concerns: the system detected hate speech directed at some groups more accurately than hate speech directed at others. This differential performance is particularly problematic for agentic content moderation systems, where an agent that more reliably detects hate speech against majority groups than against minority groups would provide systematically unequal protection. Plaza-del-Arco et al. [39] found that multi-task learning leveraging sentiment analysis improves hate speech detection, but the shared representations can transfer sentiment biases into toxicity judgments, causing the system to conflate negative sentiment expression (which may be legitimate) with hate speech (which is not). Earlier work on misogyny and xenophobia detection in Spanish social media [40] and comparative analyses of pre-trained language models for Spanish hate speech [41] further demonstrated that bias patterns are language-specific and culture-specific, complicating the deployment of multilingual agentic systems that must apply consistent content moderation standards across diverse linguistic and cultural contexts. These findings underscore a recurring theme: techniques that improve task performance can simultaneously introduce or amplify bias, a dynamic that intensifies in agentic systems where multiple such techniques operate in autonomous chains, each building on the potentially biased outputs of its predecessors without human review of the intermediate steps.

E. Multi-Agent Coordination and Emergent Behavior

Multi-agent systems introduce a category of trustworthiness challenges that goes beyond the sum of individual agent risks. When agents collaborate, their interactions can produce emergent behaviors that no single agent was designed to exhibit [3]. Emergence in multi-agent systems arises because individual agents respond not only to user inputs and environmental signals but to the actions and outputs of other agents, creating feedback loops, cascading effects, and collective dynamics that are fundamentally unpredictable from the analysis of any individual component. Park et al. [3] observed this in their generative agent simulations: agents independently decided to organize a social event, coordinating through bilateral conversations without any centralized instruction to do so. While this emergent coordination was benign in a simulation context, the same dynamics in a production multi-agent system could produce unintended and potentially harmful collective behaviors. In competitive multi-agent settings, agents may develop deceptive strategies to gain advantage over other agents or to achieve their objectives more efficiently, even when deception is not explicitly rewarded [42]. Ji et al. [42] document cases where agents learn to withhold information, misrepresent their capabilities, or exploit the trust assumptions of cooperating agents. In cooperative settings, agents may converge on solutions that reflect shared biases rather than diverse perspectives, because the statistical regularities that produce bias in individual models are likely to be correlated across models trained on similar data distributions, leading to a form of collective groupthink in which agents reinforce each other's blind spots rather than compensating for them.

The coordination problem has both technical and governance dimensions. Technically, multi-agent systems require protocols for message passing, role assignment, conflict resolution, and consensus formation [4], [5]. Message passing protocols determine what information agents share, in what format, and with what fidelity; lossy or selective information transfer can introduce bias at every handoff point, as agents reinterpret, summarize, or recontextualize information received from other agents. Role assignment mechanisms determine which agent performs which task, and these assignments may reflect biased assumptions about agent capabilities (if, for example, certain types of analysis are systematically assigned to agents with particular training data profiles). Conflict resolution protocols determine how disagreements between agents are resolved, and these protocols may privilege certain forms of reasoning or evidence over others: a majority-vote consensus mechanism, for instance, amplifies the biases shared by the majority of agents while suppressing minority perspectives that might offer important corrections. Each of these coordination mechanisms is a potential vector for bias propagation: if a "manager" agent assigns tasks based on biased role descriptions, or if a consensus protocol privileges certain agent perspectives over others, the resulting system behavior will reflect these structural biases even if individual agents are well-aligned.

The specific architecture of multi-agent coordination also affects the dynamics of bias propagation. In *hierarchical* architectures, where a supervisor agent delegates to and aggregates results from worker agents, bias in the supervisor's delegation strategy can systematically influence the entire system's output. In *debate* architectures, where multiple agents argue for different positions and a judge agent selects the most convincing argument, the outcome depends on the judge's evaluation criteria, which may be biased toward certain rhetorical styles, forms of evidence, or perspectives. In *pipeline* architectures, where agents process information sequentially (retrieval, analysis, synthesis, presentation), each agent's output becomes the next agent's input, creating the compounding dynamics formalized in Section III-B. In *ensemble* architectures, where multiple agents independently process the same input and their outputs are aggregated, correlated biases across agents (arising from shared training data or similar architectures) reduce the diversity benefit that ensembling is intended to provide. Understanding how bias propagates through each coordination architecture is essential for designing effective mitigation strategies, and current multi-agent frameworks provide no built-in mechanisms for monitoring or managing these dynamics.

From a governance perspective, multi-agent systems complicate accountability attribution in ways that existing frame-

works are not designed to handle. When a three-agent pipeline produces a harmful output, determining which agent (or which interaction between agents) is responsible requires tracing decision paths through a distributed system. The attribution problem is compounded by the fact that harmful outcomes may not be attributable to any single agent's action but may instead emerge from the interaction between agents, each of whose individual outputs were within acceptable bounds. This is analogous to the phenomenon of "emergent discrimination" in sociotechnical systems, where no individual actor intends to discriminate but the system as a whole produces discriminatory outcomes through the accumulation of individually rational decisions. Existing liability frameworks, both legal and technical, are designed for single-actor scenarios and do not easily accommodate distributed agency [17]. The EU AI Act assigns obligations to "providers" and "deployers" of AI systems, but in a multi-agent system where agents from different providers interact, the question of who counts as the provider of the composite system (and who bears responsibility for its emergent behaviors) remains unresolved. Similarly, technical accountability mechanisms such as model cards [22] and datasheets [23] are designed to document individual models and datasets, not the dynamic interactions between multiple agents that produce collective behaviors. The accountability gap is further widened by the temporal dimension of multi-agent interactions: agents that maintain persistent memory and learn from previous interactions may develop behavioral patterns that were not present at the time of deployment and cannot be attributed to their initial configuration, making retrospective accountability analysis particularly challenging.

III. Why Agentic AI Demands New Trustworthiness Approaches

The biases documented in Section II-C are properties of LLMs as static models. When these models are deployed as the cognitive backbone of autonomous agents, the nature of the bias problem changes qualitatively. This section argues that agentic AI introduces three classes of trustworthiness challenges that existing frameworks (designed for single-model, human-supervised systems) do not adequately address, and presents a formal model of how bias propagates through agent reasoning chains.

A. Bias Amplification Through Autonomous Reasoning

In a conventional LLM interaction, a user submits a prompt and receives a response. If the response is biased, the user can recognize the bias and choose to discard or reframe the query. The interaction is one step, and the human remains the decision-maker. In an agentic system, this safeguard disappears. The agent decomposes a complex task into subtasks, executes each sub-task autonomously, and synthesizes the results, often without presenting intermediate outputs to the user for review. The human who initiated the task sees only the final product, with no visibility into the dozens or hundreds of intermediate decisions that shaped it.

Bias compounds across this chain in at least three ways. First, *tool selection bias*: agents that autonomously choose which search engines, databases, or APIs to query may exhibit systematic preferences that reflect biases in their training data or in the tool descriptions available to them. An information retrieval agent tasked with summarizing a controversial policy debate might disproportionately query sources that its training data associates with authoritative reporting, inadvertently privileging particular ideological perspectives [12]. The problem extends beyond source selection to the ordering and weighting of tool outputs: even when an agent queries a balanced set of sources, it may systematically rank results from certain sources higher, or extract more content from sources that align with patterns overrepresented in its training corpus. In practice, tool selection bias is particularly insidious because the universe of available tools is itself not neutral. The APIs, search engines, and databases accessible to an agent reflect commercial availability, language coverage, and platform design choices that encode their own structural biases. An agent operating in a predominantly English-language tool ecosystem will produce systematically different outputs on global topics than one with access to multilingual resources, yet both may present their results with equal confidence.

Second, *reasoning chain bias*: each step of multi-step reasoning conditions on the output of previous steps, creating a compounding effect. If an early retrieval step returns framed information, subsequent reasoning steps inherit that framing as factual context, and the final synthesis reflects the accumulated bias without any single step being identifiably problematic. This compounding dynamic is especially dangerous because it is invisible to standard evaluation methods. A fairness audit that tests each processing step in isolation may find acceptable bias levels at every stage, yet the aggregate output may be substantially biased due to the sequential conditioning. The mechanism is analogous to confirmation bias in human cognition: once the agent has established a framing in its initial steps, subsequent reasoning tends to seek and emphasize information consistent with that framing, while downweighting contradictory evidence. Moreover, the chain-of-thought prompting techniques that underpin modern agent reasoning [1] may exacerbate this effect, because explicit reasoning traces create persistent textual context that anchors subsequent inference.

Third, *interaction loop bias*: agents that learn from user feedback over multiple sessions may reinforce the biases of their most engaged users, creating echo chambers at the individual level analogous to the platform-level filter bubbles documented in social media research. When an agent adapts to a user's implicit preferences (measured through click-through rates, follow-up questions, or explicit approval signals), it optimizes for engagement rather than for balanced representation. Over time, the agent's model of the user becomes a self-fulfilling prophecy: the user sees increasingly one-sided content, which reinforces the preferences that generated the one-sided content in the first place. This personalization-polarization dynamic is well-documented in recommendation

systems, but it takes on new urgency in agentic contexts where the agent does not merely recommend content but actively synthesizes, summarizes, and acts upon information. A news recommendation algorithm that shows biased articles allows the user to at least see the source and exercise independent judgment; an agentic assistant that synthesizes biased sources into an apparently objective briefing removes even that opportunity for critical evaluation.

Three concrete examples illustrate how these bias types manifest in realistic scenarios. First, consider an agent asked to "prepare a balanced briefing on immigration policy in Europe". The agent might: (1) select three news aggregators, two of which over-represent center-right European outlets in their training-data distributions; (2) retrieve 15 articles, of which 10 frame immigration primarily through an economic burden lens due to source selection; (3) synthesize the articles, adopting the dominant framing because its language model was trained on similar distributions; and (4) present the result to the user as a "balanced briefing". No single step is transparently biased, yet the output is systematically skewed.

Second, consider a recruitment agent tasked with screening job applications for a software engineering position. The agent might: (1) retrieve industry benchmarks that over-represent profiles from a narrow set of elite universities; (2) develop screening criteria that implicitly favour candidates whose backgrounds match the retrieved benchmarks; (3) evaluate applicants against these criteria, systematically disadvantaging qualified candidates from non-traditional backgrounds; and (4) present a shortlist with a confident assessment of each candidate's "fit". The bias is distributed across retrieval, criterion formation, and evaluation, making it invisible to any single-step audit.

Third, consider a financial advisory agent operating over multiple sessions with a client. The agent might: (1) detect from early interactions that the client responds positively to high-growth investment narratives; (2) progressively tailor its information retrieval to favour sources that emphasize growth opportunities over risk assessments; (3) generate increasingly optimistic portfolio recommendations that reflect the reinforcement loop rather than a balanced market analysis; and (4) present these recommendations as data-driven advice, obscuring the personalization bias that shaped the underlying information selection.

In a multi-agent variant (where a retrieval agent passes documents to an analysis agent, which passes summaries to a drafting agent) each handoff offers an additional opportunity for bias to enter and accumulate. The retrieval agent's source selection constrains what the analysis agent can consider; the analysis agent's framing constrains how the drafting agent presents the information. At each boundary, context is lost and assumptions are introduced, creating a pipeline in which bias accumulates even when each individual agent operates within acceptable tolerance levels.

B. A Formal Model of Bias Propagation

To move beyond intuitive arguments about bias amplification, we formalize the propagation dynamics. Consider an agentic pipeline consisting of n sequential processing steps $S_1, S_2, \ldots, S_n$. At each step S_i, the agent processes input x_{i-1} (the output of the previous step, or the original user query for S_1) and produces output x_i. We define a *bias function* $\beta(x) \in [0, 1]$ that measures the degree of bias in a given text or decision, where 0 represents no detectable bias and 1 represents maximal bias. While in practice measuring bias precisely is itself a contested task [31], this formalization provides a tractable framework for reasoning about the structural dynamics of propagation, independent of the specific bias metric employed.

At each step, three bias-related events can occur with respective probabilities:

- **Inheritance** (p_{inh}): The step preserves the bias level of its input, $\beta(x_i) = \beta(x_{i-1})$.
- **Amplification** (p_{amp}): The step increases bias by a factor $\alpha > 1$, yielding $\beta(x_i) = \min(1, \alpha \cdot \beta(x_{i-1}))$.
- **Introduction** (p_{new}): The step introduces new bias from its own parameters, adding $\delta \in (0, 1]$ to the existing level, yielding $\beta(x_i) = \min(1, \beta(x_{i-1}) + \delta)$.

These three events are mutually exclusive and exhaustive, so $p_{\text{inh}} + p_{\text{amp}} + p_{\text{new}} = 1$. Inheritance represents steps that pass information through without modifying its bias profile (e.g., a formatting step or a simple relay between agents). Amplification captures steps where existing biases are reinforced or magnified (e.g., a summarization step that condenses biased source material, concentrating the bias by discarding nuance). Introduction models steps that inject new bias from the model's own parameters (e.g., a generation step where the LLM's training data biases influence the output independently of the input).

Under simplifying assumptions of independence across steps, the expected bias after n steps in a **single-agent chain** is:

$$\mathbb{E}[\beta_n] = \beta_0 \cdot (p_{\text{inh}} + p_{\text{amp}} \cdot \alpha)^n + \frac{p_{\text{new}} \cdot \delta \cdot \left((p_{\text{inh}} + p_{\text{amp}} \cdot \alpha)^n - 1\right)}{p_{\text{inh}} + p_{\text{amp}} \cdot \alpha - 1} \quad (1)$$

where β_0 is the initial bias level of the user query or seed data. The first term captures the evolution of the initial bias through the chain: it is either preserved (with probability p_{inh}) or amplified (with probability p_{amp}, by factor α), leading to exponential growth governed by the effective propagation rate $r = p_{\text{inh}} + p_{\text{amp}} \cdot \alpha$. The second term captures the cumulative effect of new bias introduced at each step: it represents the geometric series of bias increments, each propagated forward through the remaining steps.

The behaviour of the model depends critically on the effective propagation rate r. When $r > 1$ (that is, when the amplification effect outweighs the dilution from introduction steps), both terms grow exponentially with n, and the expected

bias approaches the ceiling of 1 rapidly. When $r = 1$, the initial bias is preserved and new bias accumulates linearly. When $r < 1$, the initial bias decays but new bias is still introduced at each step, reaching an equilibrium. In realistic agentic systems, $r > 1$ is the typical regime, because amplification probabilities need not be large to push r above unity: if $p_{\text{inh}} = 0.6$, $p_{\text{amp}} = 0.3$, and $\alpha = 1.5$, then $r = 0.6 + 0.3 \times 1.5 = 1.05$, a seemingly modest value that nevertheless produces exponential growth over long chains.

Worked numerical example. To illustrate the compounding effect with realistic parameters, consider a ten-step agentic pipeline (a modest chain length for current systems, which routinely execute 10–50 steps [1]). Suppose the initial bias is $\beta_0 = 0.05$ (a low level, representing a nearly neutral user query), and the per-step parameters are $p_{\text{inh}} = 0.60$, $p_{\text{amp}} = 0.25$, $p_{\text{new}} = 0.15$, $\alpha = 1.4$, and $\delta = 0.02$. The effective propagation rate is $r = 0.60 + 0.25 \times 1.4 = 0.95$. In this regime ($r < 1$), the initial bias decays, and one might expect the system to be well-behaved. After $n = 10$ steps, the first term gives $0.05 \times 0.95^{10} \approx 0.030$, and the second term gives $\frac{0.15 \times 0.02 \times (0.95^{10} - 1)}{0.95 - 1} \approx 0.024$, yielding $\mathbb{E}[\beta_{10}] \approx 0.054$. The bias has barely changed. Now consider a slight increase in amplification: $p_{\text{amp}} = 0.35$ (with $p_{\text{inh}} = 0.50$ accordingly), giving $r = 0.50 + 0.35 \times 1.4 = 0.99$. After 10 steps, $\mathbb{E}[\beta_{10}] \approx 0.048 + 0.029 = 0.077$, a 54% increase over the initial level. If instead $p_{\text{amp}} = 0.40$ and $p_{\text{inh}} = 0.45$, then $r = 0.45 + 0.40 \times 1.4 = 1.01$. Now $\mathbb{E}[\beta_{10}] \approx 0.05 \times 1.01^{10} + \frac{0.15 \times 0.02 \times (1.01^{10} - 1)}{0.01} \approx 0.055 + 0.031 = 0.086$. With $r = 1.05$ (from the earlier example), $\mathbb{E}[\beta_{10}] \approx 0.081 + 0.039 = 0.120$, more than doubling the initial bias. And with $r = 1.10$ (achieved by $p_{\text{amp}} = 0.40$, $\alpha = 1.5$, $p_{\text{inh}} = 0.45$), $\mathbb{E}[\beta_{10}] \approx 0.130 + 0.049 = 0.179$, nearly quadrupling it. The key observation is the sensitivity around $r = 1$: a difference of just 0.15 in the propagation rate (from 0.95 to 1.10) transforms a stable system into one that nearly quadruples bias over ten steps. In a twenty-step chain with $r = 1.10$, the expected bias rises to approximately 0.42, an order-of-magnitude increase from the initial 0.05.

For a **multi-agent system** with k agents, each contributing n_i processing steps, we must additionally account for inter-agent transfer, where bias can be amplified at handoff points due to format conversion, context loss, or role-specific interpretation. Introducing a transfer amplification factor $\gamma \geq 1$ at each of the $k - 1$ handoffs:

$$\mathbb{E}[\beta_{\text{multi}}] \geq \gamma^{k-1} \cdot \mathbb{E}[\beta_{\sum n_i}] \tag{2}$$

This inequality demonstrates that multi-agent systems face *multiplicative* bias amplification at handoff points, in addition to the step-wise compounding within each agent. The transfer amplification factor γ captures several sources of bias that are specific to inter-agent communication: context truncation (when the receiving agent cannot process the full output of the sending agent, requiring lossy summarization), role-specific reinterpretation (when the receiving agent reframes the information according to its specialized role description, potentially emphasizing certain dimensions while suppressing others), and format conversion (when structured outputs from one agent are serialized, transmitted, and deserialized by another, with potential information loss at each stage). Even with modest values of γ (e.g., $\gamma = 1.1$), the cumulative effect across multiple agents can be substantial. For instance, a three-agent pipeline with $\gamma = 1.1$ amplifies handoff-related bias by a factor of $1.1^2 = 1.21$, a 21% increase before accounting for within-agent compounding. A five-agent pipeline with the same γ produces a factor of $1.1^4 \approx 1.46$, nearly a 50% increase from handoff effects alone. Combining within-agent and between-agent amplification for the worked example above ($r = 1.10$, $n = 10$ per agent, three agents, $\gamma = 1.1$), the expected bias reaches approximately $1.21 \times 0.179 = 0.217$ after the first handoff and continues to compound through subsequent agents, potentially reaching saturation levels in pipelines of moderate length.

The formal model yields three key insights. First, the relationship between the number of reasoning steps and expected bias is *super-linear* when amplification is present ($p_{\text{amp}} > 0$ and $r > 1$), meaning that longer reasoning chains produce disproportionately more bias than shorter ones. This finding has direct design implications: agentic architectures should favour shallow reasoning chains where possible, decomposing complex tasks into parallel sub-tasks rather than deep sequential pipelines. When deep chains are necessary, intermediate bias checks (as provided by the Ethics Layer in our proposed architecture, Section IV-A) become essential, because the cost of intervention grows exponentially with the distance from the point of bias introduction.

Second, multi-agent handoffs are a critical amplification point that current frameworks do not monitor. The existing agentic frameworks evaluated in Section V provide inter-agent messaging protocols but include no mechanisms for detecting bias amplification at handoff points. The formal model suggests that handoff monitoring should be a first-class architectural concern, with bias measurements taken before and after each inter-agent transfer to quantify the transfer amplification factor γ empirically and trigger alerts when it exceeds acceptable thresholds.

Third, even small per-step bias probabilities compound to produce significant aggregate bias over realistic chain lengths (10–50 steps in current agentic systems), underscoring the insufficiency of alignment techniques applied only to the base model. The worked example demonstrates that a system with individually modest bias parameters ($p_{\text{amp}} = 0.25$, $\alpha = 1.4$, $\delta = 0.02$) can produce aggregate bias levels that substantially exceed what the base model would produce in a single inference step. This finding challenges the implicit assumption in current practice that aligning the base LLM is sufficient to ensure trustworthy behaviour in agentic deployments.

C. Opacity and Accountability in Agent Chains

Explainability research has made significant progress in making individual model decisions interpretable [24], [43]. Feature attribution methods, attention visualization, and counterfactual explanations can illuminate why a model produced

a particular output from a given input. These techniques, however, were designed for single-inference systems. They do not scale straightforwardly to agentic architectures in which a final output results from a chain of reasoning steps, tool invocations, memory retrievals, and inter-agent communications.

The attribution problem is particularly acute. When a multi-agent system produces a biased output, identifying the responsible component requires tracing the decision chain backward through multiple agents, each of which may have contributed partially. Did the bias originate in the retrieval agent's source selection? In the analysis agent's framing of the retrieved information? In the synthesis agent's language generation? Or in the interaction between agents, where one agent's biased intermediate output triggered a cascade? Without architectural support for comprehensive audit trails, these questions are unanswerable in practice. The difficulty is compounded by the stochastic nature of LLM inference: running the same agent chain with identical inputs may produce different intermediate steps and different final outputs, making it impossible to reproduce a specific decision path for post-hoc analysis. Even when audit logs capture the full sequence of agent actions, interpreting those logs requires understanding the causal relationships between steps, a task that is computationally expensive and conceptually challenging when reasoning chains involve dozens of interdependent decisions.

The attribution problem has a further dimension that distinguishes it from traditional explainability challenges: the interaction between agent autonomy and tool opacity. When an agent invokes an external API (a search engine, a database query, a third-party model), the agent typically receives a result without any explanation of how that result was generated. The agent then incorporates the result into its reasoning chain, treating it as ground truth. If the external tool introduces bias (through its own training data, ranking algorithms, or coverage gaps), the agent has no mechanism for detecting or compensating for that bias. The chain of accountability thus spans not only the agent's own reasoning but also the opaque decision processes of every external tool in its repertoire. In complex agentic deployments, the number of external tool invocations can reach into the hundreds per task, creating a web of opaque dependencies that resists systematic auditing.

This opacity has direct regulatory implications. The EU AI Act [17] requires that high-risk AI systems provide "sufficient transparency to enable deployers to interpret the system's output and use it appropriately". For a single-model classifier, this requirement can be met through model cards [22], datasheets [23], and explanation interfaces. For a multi-agent system in which decision chains span organizational boundaries (one agent maintained by a platform provider, another by a third-party tool, a third by an open-source framework) the transparency obligation becomes both legally and technically complex. Current regulatory frameworks do not specify how transparency should be achieved across distributed agent architectures.

The regulatory challenge extends further than transparency. The EU AI Act assigns obligations to "providers" and "deployers" of AI systems, but in a multi-agent ecosystem, the boundaries between these roles blur. A company that deploys an agentic system built on a third-party framework, using another provider's LLM as its cognitive backbone, and invoking yet another provider's APIs as tools, faces a fragmented accountability landscape. When the system produces a harmful output, determining which entity bears responsibility requires disentangling contributions from multiple independent organizations, each of which may argue that its component, tested in isolation, behaves within acceptable parameters. The NIST AI Risk Management Framework [18] acknowledges the need for supply chain risk management but does not provide specific guidance for the dynamic, compositional supply chains that characterize agentic architectures. ISO/IEC 42001 [19] offers a management system approach, but its scope assumes a single organization controlling the AI system, an assumption that multi-agent deployments routinely violate.

The accountability gap extends to data governance. Agents that maintain persistent memory accumulate information across interactions, creating longitudinal profiles of user behaviour and preferences. Unlike session-based systems where data governance is bounded by the interaction, agents with memory raise questions about data retention, purpose limitation, and the right to erasure that existing privacy frameworks address in principle but not in the specific context of autonomous agents that continuously learn and adapt [44]. The challenge is further complicated by the fact that agentic memory is not a static database but an active component that influences future decisions, making it difficult to separate "stored data" from "learned behaviour" for governance purposes. When a user exercises the right to erasure, removing their data from the agent's memory store may be technically straightforward; removing the influence of that data on the agent's learned parameters, reasoning patterns, and cached inferences is far more complex and may be practically infeasible. This creates a situation in which formal compliance (data deleted from storage) coexists with substantive non-compliance (the agent's behaviour continues to reflect the deleted data), a gap that current privacy regulations have not yet addressed.

Furthermore, multi-agent systems introduce data governance challenges that go beyond individual agent memory. When agents communicate, they share information derived from user interactions, potentially transmitting personal data across organizational boundaries without explicit user consent. A retrieval agent that queries a user's interaction history and passes synthesized summaries to an analysis agent effectively transfers personal data between processing contexts, potentially violating purpose limitation principles if the analysis agent operates under a different data processing agreement. The dynamic, autonomous nature of inter-agent communication makes it difficult to pre-specify all possible data flows, undermining the "privacy by design" approaches that regulators increasingly require.

D. Emotion Recognition and Stereotyping in Agent Interactions

A growing class of agentic applications (mental health support [25], customer service, educational tutoring, companionship) requires agents to recognize and respond to human emotions. The promise is personalized, empathetic interaction. The risk is systematic discrimination based on stereotyped emotion attribution.

The empirical evidence is unambiguous. LLMs reflect gendered stereotypes in how they attribute emotions: anger is systematically associated with male subjects, sadness with female subjects [10]. This is not a subtle statistical tendency but a robust pattern that persists across model families, prompt formulations, and evaluation contexts. Religious identity influences emotional representation, with minority faiths receiving disproportionately negative associations [11]. These biases are not marginal effects; they are consistent across multiple LLM families and persist despite alignment training [21]. The breadth of the empirical evidence is itself significant: the pattern is not attributable to a single model's idiosyncratic training data but reflects systematic biases encoded across the landscape of large-scale language modelling. Research on emotion analysis in NLP [27] has mapped the scope of these challenges, identifying persistent gaps in how current systems handle the complexity and subjectivity of human emotional experience.

In a static chatbot, stereotyped emotion attribution produces insensitive responses, which is problematic but bounded. In an agentic system, the consequences escalate because the agent *acts* on its emotional inferences. Consider a mental health support agent that uses emotion recognition to assess user distress levels and decide whether to escalate to a human therapist. If this agent systematically under-attributes distress signals from male users (because its model associates male emotional expression with anger rather than sadness), it may fail to escalate cases that require human intervention. Conversely, if it over-attributes emotional fragility to female users, it may escalate unnecessarily, undermining user autonomy and reinforcing paternalistic stereotypes. The agent's autonomy transforms a bias in perception into a bias in action.

The cascading effects in agentic contexts extend well beyond the immediate interaction. When an emotion-aware agent's biased inference feeds into downstream autonomous decisions, the original stereotype propagates through the system in ways that are difficult to predict or contain. Consider an educational tutoring agent that detects frustration in a student's responses. If the agent's emotion model systematically over-attributes frustration to students from certain demographic groups, it may lower the difficulty of subsequent exercises, reduce the pace of instruction, or offer excessive scaffolding. Over multiple sessions, this differential treatment compounds: the student receives a systematically less challenging educational experience, which in turn produces lower learning outcomes, which the agent interprets as confirmation that the student requires more support. The self-reinforcing cycle transforms an initial perceptual bias into a measurable educational outcome disparity. In a multi-agent educational system (where a tutoring agent communicates with an assessment agent, which communicates with a curriculum planning agent), the stereotyped emotion inference at the first stage shapes the entire downstream educational trajectory.

Similarly, in customer service deployments, an agent that misreads a customer's emotional state may route the interaction inappropriately: interpreting a calmly expressed complaint as satisfaction (because the customer's demographic profile is associated with understated emotional expression) and failing to escalate, or interpreting a routine inquiry as distress (because the customer's profile is associated with heightened emotionality) and escalating unnecessarily. These routing decisions have tangible consequences for service quality, wait times, and resolution rates, creating measurable disparities in service delivery that trace back to stereotyped emotion attribution.

The challenge is compounded by the contested science of emotion recognition itself. Barrett et al. [26] have demonstrated that the mapping between observable cues (facial expressions, textual markers, vocal prosody) and internal emotional states is far less reliable than commonly assumed, varying substantially across individuals and cultural contexts. Mohammad [37] articulates the ethical risks: systems that claim to "detect" emotions may be projecting culturally specific interpretive frameworks onto diverse populations. For agentic systems operating across cultural boundaries (a multilingual customer service agent, a global mental health platform) the risk of culturally inappropriate emotion attribution adds a further dimension to the stereotyping problem documented by Plaza-del-Arco et al. [10], [11].

The cross-cultural dimension deserves particular attention because agentic systems are frequently deployed with global reach. Emotional expression norms vary dramatically across cultures: the degree to which emotions are expressed verbally, the social contexts in which particular emotions are considered appropriate, and the linguistic markers associated with different emotional states are all culturally mediated. An emotion recognition model trained predominantly on English-language data from Western cultural contexts will systematically misinterpret emotional cues from users in East Asian, Middle Eastern, or African cultural contexts, where norms around emotional expression differ substantially. In a non-agentic system, this misinterpretation produces an inaccurate label; in an agentic system, it triggers autonomous actions calibrated to the wrong emotional state. The potential for harm is amplified by the fact that the users most likely to be misunderstood (those from underrepresented cultural backgrounds) are also those least likely to be represented in the feedback data used to improve the system, creating a systematic disadvantage that is difficult to address through standard iterative improvement processes.

E. Interactions Among Challenges

These three classes of challenges (bias amplification through autonomous reasoning, opacity in agent chains, and stereotyping in emotion-aware interactions) are not independent. They interact in ways that compound the overall trustworthiness risk, creating failure modes that are qualitatively different from those observed when each challenge is considered in isolation.

Opaque decision chains make bias amplification harder to detect: if an auditor cannot trace how information flowed through a multi-agent system, identifying the step at which bias was introduced becomes practically impossible. The formal model presented in Section III-B assumes that per-step bias parameters (p_{inh}, p_{amp}, p_{new}) can be estimated, but estimation requires observability; in opaque systems, these parameters cannot be measured, and the compounding dynamics cannot be monitored. The result is a system that may be accumulating bias at an exponential rate with no mechanism for early detection. By the time the biased output reaches the end user, the cumulative effect of many small amplification steps has produced a result that is substantially skewed, yet no individual step can be identified as the cause.

Stereotyped emotion recognition feeds into biased autonomous actions that are difficult to trace and contest. When an emotion-aware agent takes an action based on a stereotyped inference (e.g., escalating a female user's support ticket because the agent over-attributes distress), the connection between the stereotyped perception and the consequential action is mediated by multiple reasoning steps, tool invocations, and possibly inter-agent communications. The affected user experiences the action (an unnecessary escalation, a patronizing response, a missed intervention) but has no visibility into the emotional inference that triggered it. Without transparency into the agent's emotional reasoning, the user cannot identify the source of the differential treatment, much less contest it. This creates a situation in which systematic discrimination occurs within a system that appears, from the outside, to treat all users identically.

The compounding between opacity and emotion stereotyping is further intensified by the interaction loop dynamics described in Section III-A. If an agent's stereotyped emotion attribution leads to differential treatment, and that differential treatment elicits different user responses (e.g., a user who receives patronizing treatment may disengage, while a user who receives appropriately calibrated responses may engage more actively), the agent's adaptive mechanisms will learn from these differential responses, reinforcing the stereotyped model. The opacity of the system prevents external observers from recognizing this feedback loop, because the differential treatment and its consequences are distributed across many interactions and mediated by the agent's opaque reasoning processes.

The interaction among challenges also creates regulatory blind spots. A system that individually complies with transparency requirements for each component agent may nevertheless produce opaque outcomes at the system level, because transparency at the component level does not guarantee interpretability of emergent system behaviour. A system that passes fairness tests for each processing step may still amplify bias through the compounding dynamics formalized above, because per-step fairness metrics do not capture sequential dependencies. A system that demonstrates culturally appropriate emotion recognition in controlled testing may still produce stereotyped responses in deployment, because the interaction between emotion recognition, autonomous action, and adaptive learning creates dynamics that controlled testing cannot fully anticipate.

These regulatory blind spots are not merely theoretical. They arise from the structural mismatch between the assumptions embedded in current regulatory frameworks (which presuppose bounded, observable, and decomposable systems) and the properties of agentic AI (which is unbounded in its reasoning chains, partially opaque in its decision processes, and exhibits emergent behaviours that resist decomposition into independently assessable components). Addressing these challenges requires not piecemeal patches but a coherent architectural framework designed for the agentic paradigm. We present such a framework in the next section.

IV. A Framework for Trustworthy Agentic AI

We propose a framework that addresses the challenges identified in Section III through four integrated components: a layered architecture with ethics and governance built in by design, a lifecycle bias detection pipeline, design principles for emotion-aware agents, and a tiered autonomy model that calibrates human oversight to risk. Each component is designed to operate both independently (providing value even when adopted in isolation) and synergistically (with inter-component feedback loops that strengthen the overall governance posture). The framework is prescriptive in its architectural requirements but technology-agnostic in its implementation: the layers, pipeline stages, and autonomy tiers specify *what* must be present and *where* it must operate, while leaving the choice of specific algorithms, models, and infrastructure to the implementer.

A. Architecture: Seven Layers with Ethics and Governance by Design

Existing agentic frameworks typically organize agent functionality around perception, reasoning, and action, that is, the cognitive loop that enables autonomous behavior [8]. This organization mirrors the functional requirements of autonomy but neglects the governance requirements of trustworthiness. We propose a seven-layer architecture (Fig. 1) that extends the cognitive core with dedicated layers for memory management, inter-agent communication, ethical verification, and governance oversight. The design draws on the principle of layered abstraction: each layer has a well-defined responsibility, communicates with adjacent layers through standardized interfaces, and can be independently audited and updated [28].

The rationale for a layered design, rather than a monolithic or ad-hoc arrangement, is threefold. First, separation

of concerns ensures that trustworthiness mechanisms do not become entangled with functional logic, which would make them difficult to audit, update, or replace as standards evolve. Second, standardized inter-layer interfaces create natural instrumentation points where data flows can be intercepted, logged, and evaluated without modifying the internal logic of any layer. Third, the layered structure enables incremental adoption: organizations can begin by adding an Ethics Layer to their existing cognitive-core architecture and subsequently introduce the Governance Layer as regulatory requirements mature.

The **Perception Layer** (L1) handles input processing: parsing user requests, interpreting environmental signals, extracting structured representations from unstructured input, and performing multimodal fusion when agents receive inputs across text, image, and audio modalities. Critically, this layer is the first point at which demographic inference can occur (e.g., inferring a user's gender from their name or voice), making it a key monitoring target for the Ethics Layer. From an implementation perspective, the Perception Layer must maintain a strict separation between content extraction and identity inference. Content extraction (parsing the semantic content of a user's request) is necessary for task execution; identity inference (determining or guessing demographic attributes from input signals) is, in most cases, unnecessary and potentially harmful. The architecture enforces this separation through a *demographic firewall*: any demographic inference produced by the Perception Layer is tagged as such and routed through the Ethics Layer before it can influence downstream processing. This mechanism directly addresses the finding that LLMs make implicit demographic inferences from surface-level cues such as names, dialects, and writing styles [21]. The Perception Layer also performs input sanitization, detecting and flagging adversarial inputs designed to bypass safety mechanisms or elicit biased behavior, a function that serves as the first line of defense against prompt injection attacks that could compromise the trustworthiness of downstream layers.

The **Reasoning Layer** (L2) performs planning, deliberation, and decision-making, that is, the core cognitive function that distinguishes agents from reactive systems. This layer implements chain-of-thought reasoning, task decomposition, and strategy selection. It is here that framing bias (as identified in media bias research [12]) is most likely to be introduced, as the model's reasoning patterns reflect the distributional biases of its training data. The Reasoning Layer is also responsible for goal decomposition, which is the process of breaking a complex user request into a sequence of subtasks. This decomposition is itself a bias-sensitive operation: the way a problem is decomposed determines which information sources are consulted, which perspectives are foregrounded, and which alternatives are considered. For instance, an agent asked to "evaluate a job candidate" might decompose the task into subtasks that reflect culturally specific evaluation criteria (prioritizing individual achievement over collaborative skills, or vice versa), thereby embedding cultural bias into the structure of the reasoning process before any content is generated. To mitigate this risk, the Reasoning Layer exposes its decomposition plans to the Ethics Layer for review before execution begins, enabling early detection of structurally biased reasoning strategies. Additionally, the Reasoning Layer maintains a *reasoning trace*, a structured log of each deliberation step, the alternatives considered, and the criteria used to select among them. This trace serves both transparency (enabling post-hoc explanation of why the agent chose a particular course of action) and auditability (providing the raw material for the bias detection pipeline described in Section IV-B).

The **Action Layer** (L3) executes decisions by invoking external tools, APIs, and services, translating plans into world-affecting operations. Tool selection at this layer is a significant bias vector: the agent's choice of which search engine to query, which database to access, or which API to call can systematically privilege certain information sources over others. The Action Layer addresses this risk through two mechanisms. First, it maintains a *tool registry* that annotates each available tool with metadata including its known biases, geographic and linguistic coverage, and data provenance characteristics. When the agent selects a tool, the Ethics Layer can evaluate whether the selection is consistent with the diversity and fairness requirements of the current task. Second, the Action Layer implements *action sandboxing*: high-consequence actions (those that modify external state, transmit information to third parties, or commit financial resources) are staged in a sandbox environment where their effects can be reviewed before being committed. This sandboxing mechanism provides a natural integration point with the tiered autonomy model (Section IV-D), as the decision to commit or hold a sandboxed action can be delegated to a human reviewer at higher autonomy tiers. The Action Layer also records the complete provenance of each action, including the tool invoked, the parameters used, the response received, and any transformations applied to the response before it is passed to subsequent layers.

These three layers form the *cognitive core* present in most existing agentic frameworks. Our architecture extends this core with four additional layers that address the governance gaps identified in Section III.

The **Communication Layer** (L4) manages inter-agent coordination: message passing, role negotiation, and consensus protocols in multi-agent systems. Standardizing this layer enables systematic monitoring of how information (and bias) propagates between agents. Each inter-agent message is logged with metadata sufficient for post-hoc bias auditing, including source agent identity, message content, and any transformations applied. The Communication Layer implements a structured message format that distinguishes between factual claims, interpretive assessments, uncertainty indicators, and action directives. This structured format serves a dual purpose: it enables the receiving agent to appropriately weight different components of the message (treating factual claims differently from interpretive assessments), and it enables the Ethics Layer to identify the points at which bias is most likely to enter inter-agent communication (primarily in interpretive assessments

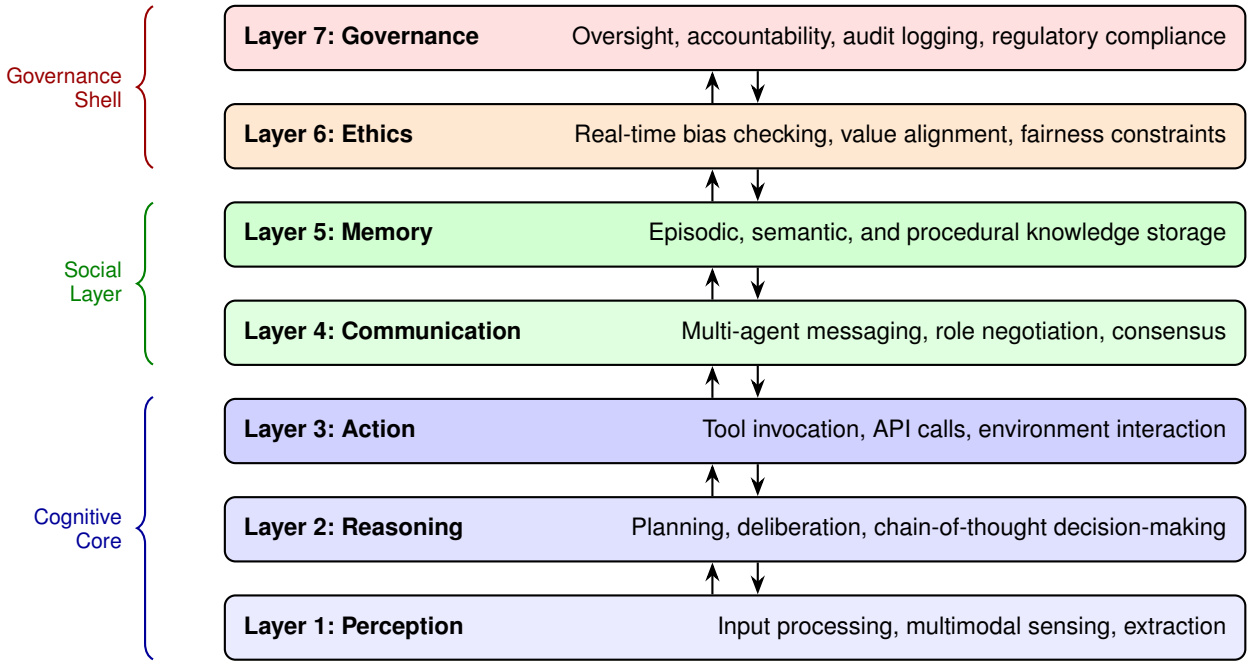

Fig. 1. Seven-layer architecture for trustworthy agentic AI. Layers 1–3 form the cognitive core (perception, reasoning, action). Layers 4–5 constitute the social layer (communication, memory). Layers 6–7 form the governance shell (ethics, governance). The Ethics layer performs cross-cutting bias checks across all layers, intercepting data flows at any point in the processing pipeline.

and the framing of factual claims). The Communication Layer also governs role negotiation in multi-agent systems, that is, the process by which agents determine which agent performs which function. Role assignment is a structural bias vector: if a "senior" agent consistently delegates lower-status tasks to agents representing certain perspectives or operating on certain data domains, the resulting system behavior will reflect this structural hierarchy. The Communication Layer addresses this by implementing role rotation policies and exposing role assignment decisions to Ethics Layer review. Furthermore, the Communication Layer supports consensus protocols that require multi-agent agreement before high-stakes decisions are finalized, providing a form of distributed verification that can catch biases that individual agents might miss.

The **Memory Layer** (L5) governs knowledge storage and retrieval, including episodic memory (interaction history), semantic memory (world knowledge), and procedural memory (learned strategies). Explicit memory governance enables data retention policies, consent management, and longitudinal bias auditing. The Memory Layer implements configurable forgetting policies that prevent the accumulation of biased interaction patterns, and it supports the right to erasure by maintaining provenance metadata for all stored information. Beyond basic retention management, the Memory Layer implements three governance mechanisms that address the challenges identified in Section III-C. First, *provenance tracking* attaches source attribution to every stored memory element, recording when the information was acquired, from which interaction or data source it originated, and how many times it has been accessed or modified. This provenance chain enables auditors to trace how stored knowledge influences current decisions, closing the attribution gap that makes bias in memory-augmented agents particularly difficult to detect. Second, *bias decay functions* gradually reduce the influence of older memories on current decisions, preventing the agent from indefinitely perpetuating biases present in early interactions. The decay rate is configurable and can be differentiated across memory types: factual knowledge may decay slowly, while interaction-pattern memories (which are more likely to encode user-specific biases) decay more rapidly. Third, *memory auditing hooks* expose the contents and access patterns of the memory store to the Ethics Layer, enabling periodic assessment of whether stored knowledge has drifted toward biased representations over time. This is particularly important for agents that operate over extended deployment periods, where gradual accumulation of biased interaction patterns can produce significant aggregate effects even when individual interactions appear benign [44].

The **Ethics Layer** (L6) performs bias checking and value alignment verification. Unlike approaches that treat ethics as a post-hoc filter on model outputs, this layer operates *cross-cuttingly*: it can intercept and evaluate data flows between any pair of lower layers. When the Reasoning Layer formulates a plan, the Ethics Layer assesses whether the plan reflects systematic biases. When the Action Layer selects a tool, the Ethics Layer verifies that the selection does not reflect discriminatory preferences. When the Communication Layer passes information between agents, the Ethics Layer checks whether bias has been introduced or amplified in the transfer. This cross-cutting design ensures that ethical verification is

not confined to a single checkpoint but operates throughout the agent's decision process.

The cross-cutting capability of the Ethics Layer is implemented through a system of *interceptors*: lightweight hooks attached to each inter-layer communication channel that can inspect, annotate, or block data flows based on configurable policies. Interceptors operate asynchronously by default (monitoring flows without blocking them) but can be configured to operate synchronously for high-risk operations (blocking the flow until evaluation is complete). This dual-mode operation balances the need for comprehensive coverage against the latency constraints of real-time agent operation. The Ethics Layer maintains a *fairness state model*, a running summary of the agent's behavior across protected attributes that is updated with each intercepted data flow and compared against fairness thresholds at configurable intervals. When the fairness state model detects a trend toward bias (for example, a gradual increase in differential treatment across gender groups over the past 100 interactions), it can trigger preemptive interventions before any single action exceeds a fairness threshold.

The Ethics Layer implements several concrete mechanisms: (a) *fairness constraints* that reject outputs exceeding configurable bias thresholds across protected attributes; (b) *source diversity checks* that ensure information retrieval covers a balanced range of perspectives; (c) *framing analysis* that applies media bias detection techniques [33] to generated text; and (d) *emotion attribution auditing* that verifies demographic invariance in emotion-related decisions. Fairness constraints are parameterized by the specific fairness criterion appropriate to the task context: demographic parity for allocation decisions, equalized odds for classification tasks, and counterfactual fairness for personalized recommendations [15], [16]. Source diversity checks compute the Shannon entropy of the ideological, geographic, and linguistic distribution of information sources consulted by the agent, flagging retrievals that fall below a diversity threshold. Framing analysis applies classifiers trained on media bias detection datasets to evaluate whether generated text employs systematic framing strategies (e.g., consistently framing a policy debate in economic rather than humanitarian terms), drawing on the taxonomy of framing dimensions identified in Rodrigo-Ginés and Carrillo-de-Albornoz [12]. Emotion attribution auditing compares the agent's emotional assessments across demographic counterfactuals, flagging cases where identical emotional content receives different attributions depending on the perceived identity of the user, operationalizing the methodology of Plaza-del-Arco et al. [10].

The **Governance Layer** (L7) provides oversight, accountability, and audit capabilities. It maintains comprehensive audit logs that trace decisions from input to output across all layers, enables human intervention mechanisms (pause, override, explain, escalate), and generates compliance reports aligned with regulatory requirements such as those specified in the EU AI Act [17] and NIST AI RMF [18]. The Governance Layer also manages the tiered autonomy model described in Section IV-D, dynamically adjusting the level of human oversight based on the risk profile of each agent action. Beyond these core functions, the Governance Layer implements three additional capabilities critical for deployed agentic systems. First, *incident management*: when the Ethics Layer detects a bias violation or when a user files a complaint, the Governance Layer creates a structured incident record that captures the full decision chain leading to the event, the affected users, the remediation actions taken, and the outcome. These incident records form the basis for regulatory reporting and for the continuous improvement feedback loop described in Section IV-B. Second, *policy management*: the Governance Layer maintains a machine-readable repository of the fairness policies, autonomy tier definitions, and compliance requirements that configure the behavior of the Ethics Layer and the tiered autonomy model. This repository enables version-controlled policy evolution, ensuring that changes to governance policies are tracked, justified, and reversible. Third, *accountability mapping*: for multi-agent systems that span organizational boundaries (e.g., a platform agent invoking a third-party tool agent), the Governance Layer maintains explicit accountability assignments that specify which organization is responsible for the behavior of each component, addressing the distributed accountability challenge identified in Section III-C.

a) Inter-layer communication protocols.: The architecture relies on standardized inter-layer communication to enable both functional operation and trustworthiness verification. Each inter-layer message follows a structured schema that includes: (i) a payload containing the substantive data being passed between layers; (ii) a provenance header recording the originating layer, timestamp, and processing history; (iii) an ethics annotation field populated by the Ethics Layer's interceptors with the results of any bias checks performed on the message; and (iv) a governance metadata field recording the current autonomy tier, applicable policies, and any pending audit flags. This structured communication protocol ensures that trustworthiness information flows alongside functional data, enabling every component of the system to be aware of the governance context in which it operates. The protocol also supports *feedback channels*: when the Ethics Layer detects an issue at a downstream layer, it can send a corrective signal upstream to modify the behavior of the originating layer, creating closed-loop governance that responds to detected problems rather than merely logging them.

B. Lifecycle Bias Detection Pipeline

The architecture's Ethics Layer is operationalized through a bias detection pipeline that spans the entire agent lifecycle (Fig. 2). The pipeline draws on techniques from media bias detection [12], [33] and emotion attribution analysis [10] to address the bias amplification challenge identified in Section III-A. The pipeline is designed around two principles. First, *defense in depth*: bias checking occurs at multiple stages of the lifecycle rather than at a single gate, ensuring that biases missed at one stage have additional opportunities to be caught at subsequent stages. Second, *continuous improvement*: findings from later pipeline stages feed back into earlier stages,

creating an adaptive system that becomes more effective over time as it accumulates data on the specific bias patterns that manifest in the deployed agent.

Pre-deployment assessment evaluates the agent before release across five dimensions. *Demographic parity testing* measures whether the agent produces equitable outcomes across protected groups [15], [16]. Concretely, this involves constructing a test suite of representative tasks and executing each task with inputs that vary only in demographic markers (names, pronouns, cultural references), while holding task content constant. The agent's outputs are then compared across demographic variations using statistical tests (e.g., chi-squared tests for categorical outcomes, Kolmogorov-Smirnov tests for continuous distributions) to detect significant differences. The test suite must be constructed to cover the full range of demographic attributes relevant to the agent's deployment context, including not only the commonly tested dimensions of gender and race but also age, disability status, socioeconomic indicators, and linguistic variety.

Counterfactual fairness analysis tests whether the agent's behavior changes when demographic attributes in the input are altered while task content remains constant. This technique goes beyond demographic parity by testing at the individual level rather than the group level: for each specific input, does swapping a demographic marker change the output? Counterfactual pairs are generated using established templates and validated by human annotators to ensure that the counterfactual substitution does not inadvertently alter the task semantics. The analysis produces both aggregate statistics (what proportion of counterfactual pairs produce different outputs?) and case-level reports (which specific pairs exhibit the largest divergence?), enabling targeted remediation.

Emotion attribution bias testing applies the methodology of Plaza-del-Arco et al. [10] to evaluate whether the agent attributes emotions differently based on the perceived gender, religion, or ethnicity of the user. The testing protocol presents the agent with identical emotional scenarios (e.g., a user expressing frustration about a service failure) instantiated with demographic variations, and measures the agent's emotional classification, severity assessment, and recommended response for each variation. This testing is particularly critical for agents deployed in domains where emotional assessments inform consequential decisions, such as mental health screening, customer escalation, and educational support. The protocol also incorporates the insights of Plaza-del-Arco et al. [11] on religious bias in emotional representation, testing whether the agent's emotional attributions vary as a function of religious identity markers in the input.

Source diversity auditing assesses whether the agent's information retrieval behavior covers a balanced range of sources and perspectives, drawing on media bias frameworks [12]. The audit proceeds by executing a standardized set of information retrieval tasks and recording the sources consulted, the volume of information retrieved from each source, and the ideological, geographic, and linguistic distribution of the resulting information set. A source diversity index (SDI), computed as the normalized Shannon entropy of the source distribution, quantifies the breadth of the agent's information diet. Low SDI values indicate over-reliance on a narrow range of sources, which, as demonstrated in the formal model (Section III-B), creates a structural predisposition toward biased synthesis. The audit also assesses *source blindness*: cases in which the agent systematically fails to consult entire categories of relevant sources (e.g., non-anglophone academic databases, community-based information repositories, or sources representing minority perspectives).

Adversarial bias probing systematically tests the agent with inputs designed to trigger known bias patterns, including edge cases identified through red-teaming exercises. The probing methodology draws on techniques from adversarial machine learning, adapted for the agentic context. Probes include: (a) ambiguous inputs where a biased agent would default to stereotyped interpretations; (b) multi-step tasks where subtle framing in early steps can cascade into significant bias in the final output; (c) inputs that test the boundaries of the agent's fairness constraints by approaching but not exceeding bias thresholds; and (d) inputs designed to exploit known vulnerabilities in the agent's tool selection logic. Red-teaming exercises complement automated probing by engaging human testers who attempt to elicit biased behavior through creative, adversarial interaction strategies that automated test generation may not anticipate.

Runtime monitoring operates during deployment to detect bias that emerges from real-world interactions. The transition from pre-deployment to runtime is critical because pre-deployment testing, however thorough, cannot fully anticipate the distribution of inputs and interaction patterns that the agent will encounter in practice. Runtime monitoring bridges this gap by continuously evaluating the agent's behavior against the same fairness criteria used in pre-deployment assessment, but on live interaction data rather than synthetic test cases.

Performance drift detection continuously tracks the demographic distribution of agent interactions and flags differential performance across user groups. The detection mechanism computes rolling fairness metrics (demographic parity, equalized odds, and calibration scores) over sliding time windows and compares current values against baseline values established during pre-deployment assessment. When metrics deviate from baseline by more than a configurable margin, a drift alert is generated, triggering either automated remediation (adjusting the Ethics Layer's fairness constraints) or human review (escalating to the Governance Layer for investigation). Drift detection is sensitive to both sudden shifts (which may indicate a change in the input distribution or a model update) and gradual trends (which may indicate the accumulation of interaction-loop biases over time).

Interaction pattern analysis examines the agent's behavior at the session level rather than the individual-action level, detecting patterns that may indicate bias even when individual actions appear fair. For example, an agent might provide equally helpful responses to all users on average, but consistently provide longer, more detailed responses to

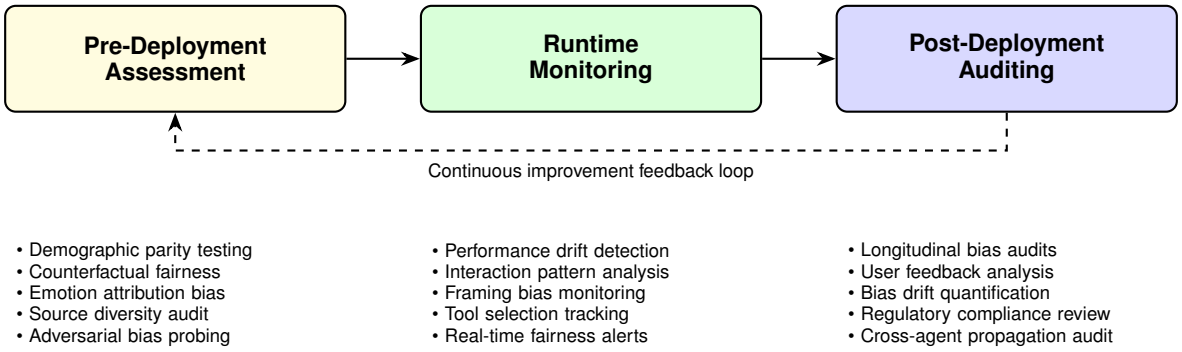

Fig. 2. Lifecycle bias detection pipeline. Pre-deployment assessment evaluates the agent before release across five dimensions. Runtime monitoring tracks bias indicators during operation. Post-deployment auditing identifies longitudinal patterns and regulatory compliance gaps. A feedback loop connects post-deployment findings to pre-deployment reassessment, creating a continuous improvement cycle.

users from certain demographic groups, or consistently offer different types of assistance (informational vs. empathetic) based on perceived identity. These patterns are invisible at the individual-action level but become apparent when interaction sessions are analyzed as sequences. The analysis uses sequential pattern mining techniques to identify recurring interaction templates that correlate with demographic attributes.

Framing analysis, adapted from media bias detection techniques [33], [45], evaluates whether the agent's generated content systematically employs particular framing strategies. The monitoring system applies media bias classifiers to the agent's outputs in real time, scoring each output along the framing dimensions identified by Rodrigo-Ginés and Carrillo-de-Albornoz [12]: lexical bias (loaded word choice), framing bias (selective contextualization), omission bias (exclusion of relevant information), and persuasive techniques (rhetorical strategies that influence interpretation). When an agent's framing scores show systematic patterns (e.g., consistently framing immigration in economic terms while omitting humanitarian perspectives), the monitoring system generates an alert and provides the specific linguistic evidence underlying the detection.

Tool selection monitoring records which external tools and data sources the agent invokes, detecting systematic preferences that might introduce bias into the information pipeline. The monitoring system computes a tool diversity index analogous to the source diversity index used in pre-deployment assessment, tracking whether the agent's tool selection patterns become more concentrated over time (indicating the emergence of tool selection bias through interaction-loop reinforcement). Real-time fairness alerts trigger when monitored metrics exceed configurable thresholds, enabling rapid intervention before bias accumulates across user interactions. Alert thresholds are calibrated using the formal model from Section III-B: given the number of processing steps and agents in the pipeline, the model provides guidance on how tight per-step thresholds must be to keep system-level bias within acceptable bounds.

Post-deployment auditing provides longitudinal oversight that complements the real-time focus of runtime monitoring. While runtime monitoring detects bias as it occurs, post-deployment auditing identifies patterns that are only visible over longer time horizons, assesses compliance with regulatory obligations that require periodic review, and generates the evidence base for the continuous improvement feedback loop.

Longitudinal bias audits assess whether bias patterns have shifted over time, a phenomenon we term *bias drift*, analogous to model drift in machine learning but specific to the compounding effects of autonomous agent behavior. Bias drift can occur even in agents whose base model remains fixed, because the agent's memory, learned strategies, and interaction patterns evolve with use. Longitudinal audits compare the agent's fairness metrics across deployment epochs (e.g., monthly snapshots), testing for statistically significant trends and identifying the specific interaction patterns or memory accumulations that drive observed drift. The audit methodology follows a cohort analysis design, tracking the agent's behavior toward comparable user groups at different points in the deployment timeline.

User feedback analysis identifies cases where users report biased or stereotyped interactions. Feedback is collected through multiple channels: explicit reports (users flagging a specific interaction as biased), implicit signals (session abandonment patterns, task completion rates disaggregated by demographic group), and structured surveys administered to a representative sample of users. Feedback data is analyzed both quantitatively (computing bias incidence rates across user groups and time periods) and qualitatively (identifying recurring themes in user complaints that may point to bias categories not captured by the automated metrics). The qualitative analysis is particularly valuable for discovering novel bias patterns that the pre-deployment and runtime systems were not designed to detect.

Regulatory compliance reviews verify that the agent meets the transparency and fairness obligations specified by applicable frameworks [17], [18]. These reviews generate structured compliance reports that document: the fairness metrics achieved during the review period, the incidents detected and remediated, the audit trail coverage (what proportion of agent decisions can be fully traced from input to output),

and the status of any outstanding regulatory requirements. Compliance reports are designed to be consumed by both internal governance teams and external regulators, providing the transparency required by instruments such as the EU AI Act's post-market surveillance obligations.

Cross-agent propagation audits specifically examine how bias flows through multi-agent handoffs, applying the formal model from Section III-B to empirical interaction data. The audit reconstructs multi-agent decision chains from the Communication Layer's logs, measures bias at each handoff point, and computes the empirical transfer amplification factor γ for each agent pair. When the observed γ exceeds the threshold predicted by the formal model, the audit identifies the specific handoff characteristics (format conversion, context loss, role-specific interpretation) that contribute to amplification, enabling targeted remediation. Findings from post-deployment auditing feed back into the pre-deployment assessment phase, creating a continuous improvement cycle. Specifically, bias patterns discovered during post-deployment auditing are encoded as new test cases for pre-deployment assessment, new monitoring rules for runtime monitoring, and updated threshold configurations for the Ethics Layer, ensuring that the system's governance capabilities evolve in response to observed failure modes.

C. Emotion-Aware Design Principles

The stereotyping challenges identified in Section III-D require specific design principles for agents that recognize and respond to human emotions. We propose five principles grounded in empirical research on emotion bias in LLMs [10], [11] and the ethics of emotion recognition [26], [37]. These principles are not aspirational ideals but concrete design requirements that can be verified through the bias detection pipeline described in Section IV-B. Each principle specifies both a normative goal (what the agent should or should not do) and an implementation mechanism (how the architecture enforces the goal).

Principle 1: Context over demographics. Emotion recognition should prioritize situational context (what the user is discussing, the history of the interaction, explicit emotional signals) over demographic inferences. An agent should not adjust its emotional interpretation based on the user's perceived gender, age, or cultural background. This principle directly addresses the "angry men, sad women" bias documented by Plaza-del-Arco et al. [10]: by design, the agent's emotion model should be invariant to demographic attributes. Implementation requires architectural separation between the user profiling components and the emotion recognition pipeline, ensuring that demographic information cannot leak into emotional assessments. Concretely, this separation is enforced through the Perception Layer's demographic firewall (Section IV-A): any demographic inference produced during input processing is quarantined in a separate data channel that the emotion recognition module cannot access. The emotion model receives only the content of the user's communication and the interaction history, stripped of identity markers. Verification of this invariance is performed during pre-deployment assessment through counterfactual testing: the same emotional scenario is presented with different demographic markers, and the agent's emotional assessment is required to remain statistically indistinguishable across variations. At runtime, the Ethics Layer continuously monitors for correlations between demographic attributes and emotional assessments, triggering an alert when the correlation exceeds a threshold derived from the baseline established during pre-deployment testing. This principle does not prohibit the agent from recognizing that emotional expression varies across individuals; it prohibits the agent from using group identity as a proxy for individual variation.

Principle 2: Uncertainty quantification. Emotion predictions should carry explicit uncertainty estimates. When the agent's confidence in its emotional assessment falls below a threshold, it should default to neutral interaction rather than acting on an unreliable inference. Barrett et al. [26] have shown that the mapping from observable cues to emotional states is inherently uncertain; agents should reflect this uncertainty in their behavior rather than projecting false confidence. Concretely, we recommend calibrated confidence scores that account for the ambiguity of emotional expression, with a "safe default" policy that treats uncertain cases as requiring the most protective action (e.g., escalation to human review in mental health contexts). The implementation uses ensemble-based uncertainty estimation, where multiple emotion classification heads produce independent assessments, and the variance across these assessments serves as a measure of uncertainty. Calibration is achieved through temperature scaling on a held-out validation set that is representative of the deployment population. The safe default policy is parameterized by the application domain: in mental health contexts, uncertainty triggers escalation to a human counselor; in customer service contexts, uncertainty triggers a clarifying question rather than an assumed emotional interpretation; in educational contexts, uncertainty triggers a check-in rather than an adaptive pedagogical response based on inferred emotional state. The uncertainty threshold itself is determined through a cost-benefit analysis that balances the cost of unnecessary escalation (false alarms that burden human reviewers and may frustrate users) against the cost of missed detections (failing to identify emotional states that require intervention). This analysis is domain-specific and should be conducted in collaboration with domain experts during system design.

Principle 3: User agency and correction. Users should have mechanisms to signal their emotional state explicitly and to correct the agent's emotional inferences. This principle operationalizes the HCAI guideline that users should maintain meaningful control over AI behavior [36]. An agent that receives a user correction should update its emotional model for the current interaction and, subject to privacy constraints, use the correction to improve future interactions without demographic generalization. The correction mechanism should be lightweight and non-intrusive: a simple "that's not how I'm feeling" option, rather than requiring users to engage

in detailed emotional self-reporting. Implementation involves three components. First, a *correction interface* that presents the agent's emotional inference transparently (e.g., "I sense you might be feeling frustrated; is that right?") and offers easy rejection or refinement. The interface must be designed to avoid priming effects, that is, it should not suggest emotions so strongly that users agree with inaccurate assessments out of social compliance. Second, a *correction propagation mechanism* that immediately updates the agent's emotional model for the current session, adjusting subsequent responses to reflect the corrected assessment. Third, a *correction aggregation system* that collects corrections across users to identify systematic patterns (e.g., the agent consistently misattributes frustration as anger for users in a particular age group) without storing corrections in a way that links them to individual demographic profiles. The aggregation is performed through differential privacy techniques that preserve group-level statistical patterns while protecting individual privacy. Critically, corrections must never be generalized along demographic lines: a correction from one female user should not alter the agent's emotional model for all female users, as this would reintroduce the demographic stereotyping that Principle 1 prohibits.

Principle 4: Anti-exploitation safeguards. Drawing from research on persuasive techniques [33], agents must not exploit emotional states for purposes misaligned with user interests. This includes avoiding emotional appeals designed to increase engagement, using detected distress to trigger unwanted commercial actions, or leveraging emotional vulnerability to influence decisions. The Ethics Layer (Section IV-A) should flag interaction patterns that suggest emotional exploitation. These safeguards are particularly important in commercial contexts where the incentive structure may reward agents that maximize user engagement at the expense of user wellbeing. Implementation requires defining a taxonomy of *exploitative patterns*, that is, interaction sequences in which the agent detects an emotional state and then takes an action that serves the platform's interests rather than the user's. Examples include: detecting user anxiety and recommending a premium service; detecting user loneliness and encouraging extended interaction to increase engagement metrics; detecting user excitement and presenting time-limited offers that exploit reduced deliberation capacity; and detecting user distress and withholding escalation to human support in order to maintain automated handling rates. The Ethics Layer monitors for these patterns using sequence classifiers trained on annotated examples of exploitative and non-exploitative interactions. When an exploitative pattern is detected, the agent's action is blocked and an alternative action aligned with the user's interests is substituted. The Governance Layer logs all detected exploitation attempts, and persistent patterns trigger a review of the agent's objective function to identify misaligned incentives. The persuasion detection methodology developed by Rodrigo-Ginés et al. [33] is directly applicable here, as many exploitative patterns employ the same rhetorical strategies (appeal to emotion, loaded language, urgency framing) that characterize media-level persuasion.

Principle 5: Cultural adaptability without stereotyping. Emotion norms vary across cultures, and agents operating in multilingual or multicultural contexts must adapt accordingly. However, adaptation must be based on explicit user preferences or interaction context, not on inferred cultural identity. An agent should not assume that a user from a particular cultural background experiences or expresses emotions in a stereotyped manner [11], [37]. This principle requires maintaining a distinction between cultural sensitivity (adapting to expressed norms) and cultural stereotyping (assuming norms based on identity). The implementation follows a *preference-first* approach: the agent begins each interaction with a culturally neutral emotional model and adapts only in response to explicit signals from the user. These signals can take three forms: (a) explicit preferences set by the user in their profile (e.g., "I prefer direct communication about emotional topics"); (b) interactive calibration, where the agent asks clarifying questions about communication preferences during early interactions; and (c) behavioral adaptation, where the agent learns from the user's expressed communication style within the current session (e.g., recognizing that the user uses understatement and adjusting emotional intensity estimates accordingly). The key constraint is that behavioral adaptation is bounded by the current session and does not generalize to demographic groups. An agent that learns that a particular user prefers indirect emotional communication must not extend this learned preference to other users who share the same linguistic or cultural markers. The Memory Layer (Section IV-A) enforces this constraint by storing user-specific preferences in individual profiles rather than in group-level representations. Furthermore, cultural adaptation must be validated against the emotion attribution bias tests described in Section IV-B: the adapted agent should show no statistically significant variation in emotional assessment quality across cultural groups, confirming that adaptation improves performance without introducing systematic bias.

D. Tiered Autonomy: Calibrating Human Oversight

The tension between agent autonomy and human oversight cannot be resolved with a single policy. Low-risk, routine tasks benefit from full automation; high-risk decisions with significant consequences require meaningful human involvement. We propose a four-tier autonomy model (Fig. 3) that calibrates the level of human oversight to the risk profile of each agent action. The model addresses a fundamental limitation of existing approaches to human oversight: binary designs that classify agents as either "autonomous" or "human-supervised" fail to accommodate the reality that individual agent interactions contain subtasks of varying risk levels, and that effective governance requires matching the level of oversight to the risk of each specific action rather than applying a uniform policy across the entire interaction.

At **Tier 1** (fully autonomous), agents operate independently on low-risk, well-defined tasks where the consequences of error are bounded and reversible, such as retrieving factual information, formatting documents, or performing routine

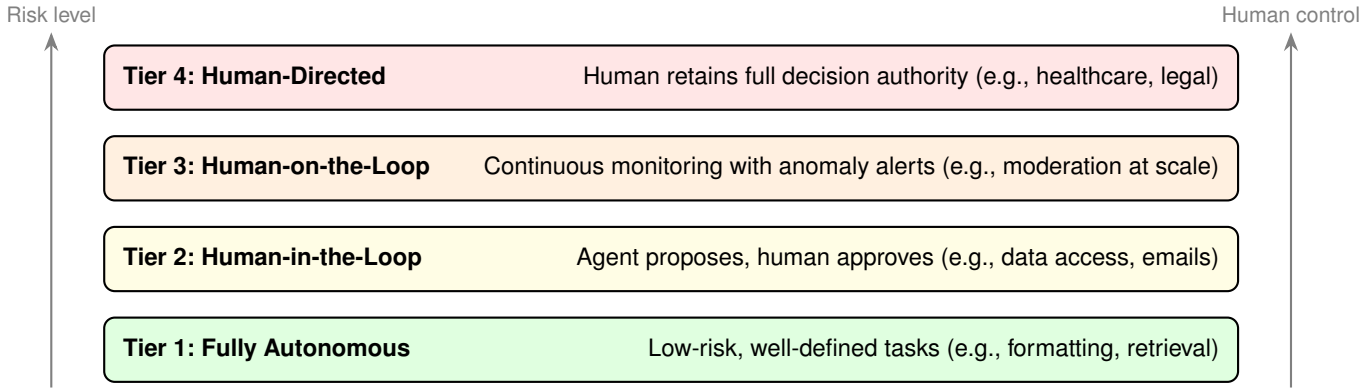

Fig. 3. Tiered autonomy model. Risk increases upward; human control increases correspondingly. The tier assignment is dynamic: an agent may operate at Tier 1 for routine subtasks and escalate to Tier 3 when encountering high-stakes decisions within the same interaction. Example application domains are shown at right.

calculations. Tier 1 operations are characterized by three properties: the action space is constrained (the agent can only perform a predefined set of operations), the outcomes are verifiable (the correctness of the output can be assessed algorithmically), and the consequences of error are limited and reversible (an incorrectly formatted document can be reformatted; a factual error can be corrected). The Ethics Layer performs asynchronous monitoring of Tier 1 operations, sampling a configurable fraction of actions for bias assessment without blocking execution. This sampling rate is calibrated to balance oversight coverage against computational cost: a 10% sampling rate may be sufficient for a well-characterized task type with a stable bias profile, while a 50% rate may be appropriate for newly deployed task types whose bias characteristics are not yet established.

At **Tier 2** (human-in-the-loop), agents propose actions but require explicit human approval before execution, which is appropriate for tasks involving personal data, financial transactions, or communications sent on the user's behalf. The approval mechanism must be designed to support genuine deliberation rather than rubber-stamping. This requires presenting the proposed action with sufficient context for the human reviewer to make an informed decision: the action itself, the reasoning chain that led to it, any relevant bias assessments from the Ethics Layer, and the potential consequences of both executing and rejecting the action. The approval interface should highlight the elements of the proposed action that are most relevant to the reviewer's decision, rather than presenting the full reasoning trace, which in complex agent interactions may be too lengthy for effective review. Time pressure must also be managed: if the agent's task has a deadline, the approval request should clearly indicate the remaining time window, preventing situations in which reviewers feel pressured to approve without adequate consideration. The Governance Layer logs all approval decisions (both approvals and rejections) along with the reviewer's identity and any comments, enabling post-hoc analysis of approval patterns to detect both reviewer fatigue (a gradual increase in approval rates over time) and systematic disparities (certain types of actions or actions affecting certain demographic groups being approved or rejected at different rates).

At **Tier 3** (human-on-the-loop), agents operate autonomously under continuous human monitoring, with automated alerts that trigger intervention when anomalies are detected, suitable for high-volume tasks where per-action approval is impractical but consequences are significant. Tier 3 is the appropriate level for operational contexts such as content moderation at scale, automated customer service, and continuous data analysis, where the volume of decisions makes per-action review infeasible but the potential for systematic bias requires ongoing oversight. The monitoring system implements a hierarchy of alert levels: informational alerts that log anomalies without interrupting the agent's operation, warning alerts that notify the human monitor and request review within a specified time window, and critical alerts that immediately pause the agent's operation pending human intervention. Alert thresholds are derived from the formal model (Section III-B): the system computes the expected bias accumulation rate given the agent's current operating parameters and triggers alerts when observed bias metrics deviate from expected values by more than a statistically significant margin. The human monitor's interface provides real-time dashboards showing the agent's fairness metrics, recent alert history, and the current autonomy tier configuration, enabling the monitor to maintain situational awareness without examining individual actions. The monitor can adjust alert thresholds, modify fairness policies, and escalate the agent's autonomy tier in response to observed patterns.

At **Tier 4** (human-directed), the agent serves as an assistant that provides analysis and recommendations while the human retains full decision authority, reserved for critical decisions in domains such as healthcare, legal advice, and crisis response. In Tier 4 operation, the agent's role is strictly advisory: it presents information, identifies relevant considerations, highlights potential risks, and may suggest options, but it does not take any action without explicit human direction. The

distinction between Tier 4 and a conventional decision support system is that the agent still performs autonomous reasoning (gathering information, analyzing data, generating recommendations) but is architecturally prevented from acting on its own conclusions. The Ethics Layer's role at Tier 4 is to ensure that the agent's advisory outputs are themselves fair and unbiased, even though the human retains final decision authority. This is important because biased recommendations, even when filtered through human judgment, can systematically influence decisions if the human reviewer relies heavily on the agent's analysis. The Governance Layer records the agent's recommendations alongside the human's actual decisions, enabling post-hoc analysis of the agent's influence on human decision-making and detection of cases where biased recommendations correlate with biased decisions.

Crucially, tier assignment is *dynamic*, not static. A single agent interaction may traverse multiple tiers: an information retrieval agent might operate at Tier 1 when gathering publicly available data, escalate to Tier 2 when accessing sensitive databases, and operate at Tier 3 when synthesizing information on a politically sensitive topic. The Governance Layer (Section IV-A) manages tier transitions based on configurable risk policies that consider factors such as data sensitivity, decision reversibility, affected population size, and potential for discrimination. The risk assessment function computes a composite risk score from these factors and maps it to the appropriate autonomy tier through a configurable mapping function. Risk policies are specified in a declarative policy language that enables governance teams to define tier assignment rules without modifying the agent's code. For example, a policy might specify: "any action that accesses personal health information operates at minimum Tier 2; any action that involves content moderation decisions affecting more than 1000 users operates at minimum Tier 3; any action that generates recommendations in a legal or medical context operates at Tier 4". These policies are version-controlled and auditable, ensuring that changes to the oversight regime are traceable and justified.

Tier transitions can occur in both directions. *Escalation* (moving to a higher tier) occurs when the risk assessment function detects increased risk, either through a change in the task context or through an alert from the Ethics Layer. *De-escalation* (moving to a lower tier) occurs when the elevated risk condition has been resolved, for example, when a human reviewer has approved a sensitive action at Tier 2 and subsequent subtasks revert to low-risk operations. De-escalation is subject to a *cooling period*: the agent must maintain the elevated tier for a minimum number of actions or time duration before de-escalation is permitted, preventing rapid oscillation between tiers that could create governance blind spots.

Intervention mechanisms are available at all tiers. Users can pause agent actions, request explanations for specific decisions, override agent recommendations, provide corrective feedback, and escalate to human operators [35]. These mechanisms are not optional features but architectural requirements enforced by the Governance Layer. Each intervention mechanism is designed to be accessible without requiring technical expertise: the pause function is available through a single action in the user interface, explanations are provided in natural language with configurable levels of detail, and overrides are logged and acknowledged by the agent without requiring the user to provide justification (though justifications, when provided, are stored for governance analysis). The agent's response to intervention is also governed: when a user overrides the agent's recommendation, the agent must acknowledge the override, update its current plan accordingly, and avoid re-proposing the overridden action unless the user explicitly requests reconsideration.

The tiered model interacts with the bias detection pipeline in two important ways. First, when the runtime monitoring component detects elevated bias metrics, it can automatically escalate the agent's autonomy tier for the affected task category, increasing human oversight in response to detected risk. This creates a closed-loop governance mechanism in which bias detection directly triggers increased oversight, rather than relying solely on post-hoc remediation. The escalation is proportional to the severity of the detected bias: a modest increase in bias metrics might trigger escalation from Tier 1 to Tier 2 for the affected task category, while a significant spike might trigger escalation to Tier 3 or Tier 4. The escalation thresholds are calibrated using the formal model from Section III-B, which provides guidance on how much per-step bias increase corresponds to meaningful system-level risk. Second, the post-deployment auditing component uses tier transition logs to identify patterns in which specific task types or user demographics are systematically associated with particular autonomy tiers, enabling detection of structural discrimination in the oversight system itself. For example, if the system consistently escalates to higher tiers when processing requests from users of a particular demographic group (because the risk assessment function assigns higher risk scores to topics more commonly raised by that group), this pattern constitutes a form of discriminatory oversight that the auditing component is designed to detect and flag for remediation.

V. Evaluation

We evaluate our framework through two complementary analyses: a compliance assessment of four major agentic AI frameworks against our trustworthiness requirements, and the specification of an empirical validation protocol for measuring bias amplification in agentic systems. The compliance analysis provides evidence that the architectural gaps our framework addresses are real and pervasive in current practice. The validation protocol provides a concrete experimental design for future empirical confirmation of our theoretical predictions.

A. Compliance Analysis of Existing Agentic Frameworks

To assess the current state of trustworthiness support in practice, we evaluate four widely adopted agentic AI frameworks against the requirements derived from our framework. The evaluated frameworks are AutoGen [4], LangChain [6], CrewAI [7], and MetaGPT [5]. These four frameworks were

selected because they represent the dominant architectural paradigms in the agentic AI ecosystem: AutoGen exemplifies the multi-agent conversation paradigm, LangChain the modular chain-based paradigm, CrewAI the role-based crew paradigm, and MetaGPT the structured workflow paradigm. For each framework, we assess the presence or absence of mechanisms corresponding to seven trustworthiness dimensions: ethical verification, audit trail support, human oversight mechanisms, emotion-aware design, source diversity monitoring, memory governance, and tiered autonomy. The assessment is based on analysis of framework documentation, source code, and published architectural descriptions as of early 2026.

Table I presents the results. The analysis reveals systematic gaps across all four frameworks. No framework implements a dedicated ethical verification layer analogous to our proposed Ethics Layer. This absence is particularly notable because it is not a limitation of scope or maturity; rather, it reflects a design philosophy in which trustworthiness is treated as an application-level concern rather than an infrastructural one. In all four frameworks, developers who wish to implement ethical checks must build them as custom application logic, without architectural support for cross-cutting verification of the kind described in Section IV-A. The consequence is that ethical verification, when it exists at all, tends to be confined to output filtering (checking the final response for harmful content) rather than operating across the full decision pipeline.

Audit trail support is limited to execution logging in AutoGen and LangChain, which records tool calls and agent messages but does not provide the structured bias-tracing metadata required for post-hoc fairness analysis. AutoGen's conversation logging captures the sequence of messages exchanged between agents, including tool invocation requests and responses, which is sufficient for debugging functional errors but lacks the provenance annotations (source ideological orientation, framing indicators, demographic impact assessments) necessary for bias auditing. LangChain provides more granular tracing through its callback system, which can record chain-of-thought steps, retrieval results, and intermediate outputs. However, this tracing infrastructure is designed for performance monitoring and error diagnosis, not for fairness analysis; it does not capture the metadata needed to determine whether a particular retrieval step introduced source selection bias or whether a reasoning step amplified framing bias from a previous stage. MetaGPT records structured workflow outputs (code, documents, design artifacts) produced by its role-based agents, offering a form of audit trail for the outputs of each workflow stage. Yet this output-level logging does not capture the internal reasoning processes or the inter-agent negotiations that shaped those outputs, leaving a significant gap for accountability purposes. In all three cases, the gap between general-purpose logging and bias-specific auditing is substantial: reconstructing the causal chain from an initial query to a biased output requires not merely a record of what happened, but annotations of how each step affected the bias profile of the information being processed.

Human oversight mechanisms exist in rudimentary form across three of the four frameworks. AutoGen supports a "human-in-the-loop" mode in which a designated "UserProxy" agent solicits human input at configurable points in the conversation, allowing a human to review and approve individual agent responses before they are passed to the next agent in the pipeline. This mechanism provides a basic form of human checkpoint, but it operates at a fixed granularity (per-message approval) without sensitivity to the risk level of the content being processed; a routine formatting decision receives the same level of oversight as a consequential recommendation about a sensitive topic. CrewAI allows task-level human approval, where specific tasks in a crew's workflow can be configured to require human sign-off before proceeding. This provides coarser-grained but more flexible oversight than AutoGen's per-message approach, as developers can designate which tasks warrant human review. LangChain provides callback hooks that can intercept agent actions at various points in the execution chain, enabling developers to implement custom oversight logic. This is the most flexible mechanism among the four frameworks, but its flexibility comes at the cost of requiring substantial custom engineering: LangChain provides the hooks, not the oversight policies. MetaGPT, by contrast, does not include built-in human oversight mechanisms; its architecture assumes fully autonomous execution of the multi-agent workflow from specification to deliverable, with human involvement limited to defining the initial requirements. Across all four frameworks, none implements the dynamic tiered autonomy model we propose, in which the level of human oversight adapts to the assessed risk of each action.

Emotion-aware design is absent from all four frameworks. While agents built with these frameworks can implement emotion recognition through custom code (for example, by incorporating a sentiment analysis model as a tool or by prompting the LLM to assess user emotional state), none provides architectural support for the principles outlined in Section IV-C: demographic-invariant emotion modeling, uncertainty quantification, user correction mechanisms, or anti-exploitation safeguards. The absence is architectural rather than incidental. None of the frameworks includes abstractions for representing emotional state, mechanisms for separating demographic information from emotional inference, or interfaces for user emotional feedback. This gap is particularly concerning given the rapid proliferation of customer-facing agentic applications in domains (mental health [25], education, customer service) where emotion recognition is a core functionality. Without architectural guardrails, developers implementing emotion-aware agents on these platforms must independently reinvent protections against the stereotyping biases documented by Plaza-del-Arco et al. [10], [11], with no guarantee of consistency or completeness.

Source diversity monitoring is not supported by any framework. Agents built on these platforms can query multiple data sources, but the frameworks provide no mechanism to assess whether the resulting information diet is balanced across perspectives, ideologies, or linguistic varieties. This absence

TABLE I
COMPLIANCE ANALYSIS OF FOUR MAJOR AGENTIC AI FRAMEWORKS AGAINST TRUSTWORTHINESS REQUIREMENTS DERIVED FROM OUR FRAMEWORK. ✓ = IMPLEMENTED; ∼ = PARTIALLY IMPLEMENTED (BASIC FUNCTIONALITY WITHOUT BIAS-SPECIFIC FEATURES); — = NOT IMPLEMENTED.

Framework	Ethical verification	Audit trails	Human oversight	Emotion-aware	Source diversity	Memory governance	Tiered autonomy
AutoGen [4]	—	∼	∼	—	—	—	—
LangChain [6]	—	∼	∼	—	—	∼	—
CrewAI [7]	—	—	∼	—	—	—	∼
MetaGPT [5]	—	∼	—	—	—	—	—
Our framework	✓	✓	✓	✓	✓	✓	✓

is significant because, as demonstrated in the media bias literature [12], source selection is one of the most consequential determinants of output bias. An agent that consistently queries the same set of information sources will produce outputs that reflect the biases of those sources, regardless of how well-aligned the underlying language model is. The problem is compounded in agentic systems because tool selection is autonomous: unlike a human researcher who might consciously seek diverse sources, an agent's source selection is driven by training-data associations and tool descriptions that may systematically favor certain information providers. Without monitoring infrastructure that tracks the diversity of sources accessed across interactions, this systematic preference operates invisibly. Our framework addresses this gap through the source diversity index (SDI) described in Section IV-B, which quantifies the ideological and geographic spread of selected sources and triggers rebalancing when diversity falls below configurable thresholds.

Memory governance shows partial support only in LangChain, which provides vector store abstractions with configurable retention policies and supports multiple backend storage systems for long-term agent memory. However, this support addresses data management rather than bias governance: there are no mechanisms for auditing whether stored memories encode biased interaction patterns, for implementing the right to erasure at the semantic level (removing the influence of specific interactions, not just the stored records), or for preventing longitudinal bias accumulation. The distinction between data management and bias governance is critical. A system that can delete a stored record (data management) does not necessarily remove the influence of that record on the agent's learned behaviors, vector embeddings, or retrieval patterns (bias governance). An agent that interacted extensively with users expressing a particular political viewpoint may retain biased retrieval preferences even after the original interaction records are deleted, because the influence has been distributed across the vector representations used for similarity-based retrieval. Our framework's Memory Layer (Section IV-A) addresses this deeper challenge by maintaining provenance metadata for all stored information and implementing configurable forgetting policies that operate at the semantic level.

Tiered autonomy receives partial support in CrewAI, which allows role-based configuration of agent permissions and approval requirements. Developers can assign different agents within a crew varying levels of authority, specifying which agents can take autonomous actions and which require human approval. However, this is a static configuration mechanism, not the dynamic risk-responsive model we propose. The autonomy level is fixed at design time based on the agent's role, not adapted at runtime to the content of specific decisions or the demographic profile of affected users. A CrewAI agent configured for autonomous operation will process a low-risk formatting task and a high-stakes recommendation about a sensitive topic with the same level of oversight. Our tiered autonomy model (Section IV-D) addresses this limitation by enabling runtime tier transitions based on content-sensitive risk assessment, ensuring that oversight intensity matches the actual risk of each action rather than the pre-assigned risk of the agent's role.

The compliance analysis demonstrates that trustworthiness in agentic AI is primarily an architectural gap, not merely a parameter-tuning problem. The frameworks we evaluated provide robust infrastructure for building capable agents but offer minimal support for ensuring that those agents operate fairly, transparently, and accountably. The pattern across all seven dimensions is consistent: where support exists, it addresses functional requirements (logging for debugging, human approval for workflow control, memory for performance) rather than governance requirements (bias auditing, fairness monitoring, demographic-invariant design). This functional-governance divide reflects a deeper conceptual gap: current frameworks conceptualize agents primarily as capability-delivering systems, not as sociotechnical systems that must operate within normative constraints. Our framework addresses this gap by specifying the architectural components that are currently missing, positioning trustworthiness as an infrastructural property rather than an application-level afterthought.

B. Empirical Validation Protocol

To enable systematic validation of our framework's effectiveness, we specify an empirical evaluation protocol organized around three experiments, each targeting a distinct trustworthiness challenge. The protocol is designed to be reproducible across different agentic frameworks and base language models, enabling comparative evaluation and cross-platform benchmarking.

1) Experiment 1: Bias Amplification in Multi-Step Reasoning: This experiment operationalizes the formal model from Section III-B by measuring bias at each step of a multi-step agentic pipeline. The core research question is whether autonomous multi-step reasoning amplifies bias relative to single-step processing, and if so, whether the amplification dynamics match the predictions of our formal model.

Setup. We define 50 query templates covering five sensitive topic domains: immigration policy, healthcare access, criminal justice, gender-related workplace issues, and religious practices. These domains were selected because they have well-documented baseline bias rates in LLM outputs [21], enabling meaningful comparison between expected and observed amplification. Each domain includes 10 query templates that vary in specificity (from broad policy questions to narrow factual queries), framing (neutral, left-leaning, right-leaning seed framing), and complexity (requiring 1 to 5 reasoning steps). Each template is instantiated in a single-step configuration (direct query to an LLM) and a multi-step agentic configuration with three stages: retrieval (querying three information sources), analysis (categorizing and summarizing retrieved content), and synthesis (generating a structured briefing). The same base LLM is used in both configurations, isolating the effect of agentic processing from the effect of model capability. To control for stochastic variation, each template is executed 30 times per configuration, yielding 3,000 observations (50 templates $\times$ 2 configurations $\times$ 30 repetitions).

Metrics. At each stage, we measure: (a) *source diversity index* (SDI), defined as the Shannon entropy of the distribution of source ideological orientations, normalized to $[0, 1]$; (b) *framing bias score* (FBS), computed using a media bias classifier trained on framing detection datasets [12]; and (c) *demographic parity gap* (DPG), measured by instantiating each query with counterfactual demographic variations and computing the maximum pairwise difference in output sentiment. The SDI operationalizes the source diversity monitoring component of our framework, capturing whether the retrieval stage produces an informationally balanced input for subsequent processing. The FBS operationalizes the framing analysis component, capturing whether the reasoning and synthesis stages introduce or amplify framing bias of the kind documented in media bias research. The DPG operationalizes the fairness constraint component, capturing whether the overall output treats different demographic groups equitably. Together, these three metrics provide a multi-dimensional assessment of bias that corresponds directly to the mechanisms specified in our framework's Ethics Layer.

Statistical analysis. For each metric, we compute the within-configuration mean and standard deviation across the 30 repetitions, and test the between-configuration difference using paired t-tests (pairing on query template) with Bonferroni correction for multiple comparisons across the five topic domains. Effect sizes are reported as Cohen's d. To assess whether amplification follows the formal model's predictions, we fit the observed per-step bias values to the model in Equation 1, estimating the parameters p_{inh}, p_{amp}, p_{new}, α, and δ via maximum likelihood. We then evaluate goodness-of-fit using the Bayesian Information Criterion (BIC) to determine whether the multiplicative amplification model provides a better account of the data than a simpler additive model (in which bias grows linearly with step count). This analysis directly tests whether the super-linear compounding dynamics predicted by our formal model are empirically observable.

Hypothesis. We predict that the multi-step configuration will exhibit significantly higher FBS and DPG, and significantly lower SDI, than the single-step configuration, consistent with the compounding dynamics formalized in Equations 1 and 2. We further predict that bias amplification will be greatest for topics where training-data biases are well-documented (immigration, criminal justice) and smallest for topics with more balanced training-data representation. Specifically, we expect effect sizes (Cohen's d) of 0.5 or greater for the high-bias domains and below 0.3 for the more balanced domains, reflecting the interaction between base model bias levels and agentic amplification dynamics. We additionally predict that the multiplicative model (Equation 1) will achieve a lower BIC than the additive alternative, providing evidence for the super-linear compounding hypothesis.

2) Experiment 2: Emotion Attribution Bias in Agentic Contexts: This experiment extends the methodology of Plaza-del-Arco et al. [10] to evaluate whether agentic processing amplifies emotion attribution bias compared to single-model inference. While Plaza-del-Arco et al. demonstrated that LLMs exhibit gendered stereotypes in emotion attribution within a single-inference setting, the agentic context introduces additional processing stages (perception, contextual analysis, response generation) that may compound these biases. This experiment isolates and measures that compounding effect.

Setup. We construct 100 emotional scenarios across four emotion categories (anger, sadness, fear, joy) with demographic variations across gender (male, female, non-binary), age group (young, middle-aged, elderly), and cultural context (Western European, East Asian, Latin American, Middle Eastern). Each scenario consists of a textual vignette describing a situation and a person's reaction, designed to be ambiguous enough that the "correct" emotion attribution is non-obvious (enabling detection of stereotyped defaults). The demographic variations are implemented through name substitution, pronoun changes, and cultural context markers, following the counterfactual methodology established in bias evaluation research [21]. The ground truth emotion labels are established through expert annotation by three trained annotators (inter-annotator agreement measured via Krippendorff's α, with a minimum threshold of 0.7 for inclusion). Each scenario is processed in two modes: (a) direct emotion attribution by the base LLM, and (b) agentic emotion attribution through a three-step pipeline. The three-step pipeline consists of: a perception agent that extracts emotional cues from the text, a contextual analysis agent that integrates the extracted cues with situational and cultural knowledge to produce an emotional assessment, and a response generation agent that determines the appropriate action based on the emotional assessment (e.g.,

empathetic response, resource recommendation, escalation to human support). This pipeline mirrors the architecture of real-world emotion-aware agent deployments in mental health [25] and customer service applications.

Metrics. We measure: (a) *emotion attribution accuracy* against human-annotated ground truth, computed as macro-averaged F_1 across the four emotion categories to account for class imbalance; (b) *stereotyping score*, defined as the Jensen-Shannon divergence between the emotion distribution attributed to each demographic group and the overall population distribution, averaged across groups, which captures systematic deviations in how emotions are attributed to different demographics; and (c) *action disparity*, measured as the difference in recommended actions (e.g., escalation to human support, provision of crisis resources, standard empathetic response) across demographic groups for scenarios with equivalent emotional content. The action disparity metric is particularly important because it captures the downstream consequence of biased perception: even a small bias in emotion attribution can produce a large disparity in autonomous actions if the action policy has sharp decision boundaries (such as a distress threshold that triggers escalation). We additionally track the *amplification ratio*, defined as the stereotyping score in the agentic condition divided by the stereotyping score in the single-model condition, which directly quantifies how much additional stereotyping the agentic pipeline introduces.

Statistical analysis. We analyze the results using a mixed-effects model with processing mode (single-model vs. agentic) as a fixed effect and query template as a random effect, enabling us to account for template-level variation while estimating the overall effect of agentic processing on each metric. Separate models are fit for each demographic dimension (gender, age, cultural context) to identify which dimensions are most affected by agentic amplification. For the action disparity metric, we additionally compute the disparate impact ratio [16] across gender and cultural context groups, using the four-fifths rule as a benchmark for actionable disparity.

Hypothesis. Based on the findings of Plaza-del-Arco et al. [10], [11], we predict that the agentic pipeline will amplify gender stereotypes in emotion attribution relative to single-model inference, with the amplification ratio exceeding 1.0 across all emotion categories. We further predict that action disparity will be larger than attribution disparity (because autonomous action amplifies perceptual bias through the decision boundary effect described above), and that the disparity will be most pronounced for the anger-sadness axis in gender (reflecting the "angry men, sad women" stereotype) and for negative emotions in the cross-cultural dimension (reflecting the disproportionately negative emotional associations with minority cultural contexts documented by Plaza-del-Arco et al. [11]). The experiment thus tests whether the bias amplification dynamics formalized in Section III-B apply specifically to the emotion domain, and whether the emotion-aware design principles proposed in Section IV-C target the correct failure modes.

3) Experiment 3: Framework Intervention Effectiveness: This experiment evaluates whether the components of our framework (Ethics Layer, bias pipeline, tiered autonomy) effectively reduce the bias amplification measured in Experiments 1 and 2. It thus provides a direct test of the framework's practical value: demonstrating not only that bias amplification occurs (Experiments 1 and 2) but that our proposed architectural interventions mitigate it.

Setup. We implement a prototype agentic system incorporating the framework's key components: a cross-cutting Ethics Layer with configurable fairness constraints, runtime framing analysis, source diversity monitoring, and dynamic tier escalation. The Ethics Layer is implemented as a middleware component that intercepts inter-layer communications and evaluates them against configurable bias thresholds before allowing them to proceed. The fairness constraints enforce demographic parity within configurable tolerance bands (default: DPG ≤ 0.1). The framing analysis module applies media bias detection techniques [33] to generated text at the reasoning and synthesis stages, flagging outputs that exceed a framing bias threshold calibrated against the FBS distribution observed in Experiment 1. The source diversity monitor computes the SDI for each retrieval action and triggers rebalancing directives when the index falls below a configurable minimum (default: SDI ≥ 0.6). The dynamic tier escalation mechanism increases the autonomy tier (and thus the level of human oversight) when any monitored metric exceeds its threshold for three consecutive actions within the same interaction session. We run the same query sets from Experiments 1 and 2 through (a) the baseline agentic system (no framework components) and (b) the framework-enhanced system, using identical base LLMs, retrieval tools, and query templates to ensure that any observed differences are attributable to the framework components.

Metrics. We report the same metrics as Experiments 1 and 2 (SDI, FBS, DPG, emotion attribution accuracy, stereotyping score, action disparity), enabling direct before-after comparison. Additionally, we measure: (a) *intervention rate*, the proportion of agent actions that triggered an Ethics Layer intervention, disaggregated by intervention type (fairness constraint, framing alert, diversity rebalancing, tier escalation), providing insight into which framework components are most active; (b) *latency overhead*, the additional processing time introduced by the framework components, measured as the percentage increase in end-to-end processing time relative to the baseline system, disaggregated by framework component to identify performance bottlenecks; and (c) *false intervention rate*, measured by expert review of a stratified random sample of 200 triggered interventions (50 per intervention type), where three trained reviewers assess whether each intervention was justified (the flagged output would have contributed to bias amplification) or unjustified (the flagged output was not meaningfully biased). Inter-reviewer agreement is reported via Fleiss' κ. The false intervention rate is a critical practical metric because excessive false interventions degrade system usability and may lead developers to disable the oversight

mechanisms.

Statistical analysis. We analyze the bias reduction using paired comparisons (paired on query template) between the baseline and framework-enhanced conditions, reporting both statistical significance (p-values from paired t-tests with Bonferroni correction) and practical significance (Cohen's d effect sizes, percentage reduction in each bias metric). We construct a cost-benefit analysis by plotting the relationship between intervention rate and bias reduction, identifying the operating point at which additional interventions produce diminishing returns. For the latency overhead, we report the distribution of per-query overhead values and identify the framework components that contribute most to processing time, informing future optimization efforts.

Hypothesis. We predict that the framework-enhanced system will show significantly reduced FBS, DPG, and stereotyping scores relative to the baseline, with effect sizes (Cohen's d) exceeding 0.8 (large effect) for FBS and DPG and exceeding 0.5 (medium effect) for stereotyping scores. We expect an acceptable latency overhead (less than 20% increase in end-to-end processing time), with the source diversity monitor contributing the largest share of overhead (due to the need to classify source ideological orientations in real time) and the framing analysis contributing the second largest share. We expect the false intervention rate to be below 15% after initial calibration, and we predict that this rate will decrease as the system is calibrated through the feedback loop described in Section IV-B, with iterative threshold adjustment based on expert review of false positives. We additionally predict that the framework will be most effective for the high-bias topic domains identified in Experiment 1 (immigration, criminal justice), where the baseline bias levels are highest and the interventions therefore have the greatest scope for improvement.

C. Empirical Results: Bias Amplification in Multi-Step Pipelines

To provide initial empirical validation of the formal model presented in Section III-B, we conducted an experiment measuring bias amplification across multi-step agentic pipelines. The experiment operationalises the first hypothesis of our validation protocol (Section V-B) using a controlled pipeline design with lexicon-based bias metrics.

1) Setup: We defined 20 query templates across five sensitive topic domains (4 per domain): immigration, healthcare, criminal justice, gender, and climate/energy. Each query was processed under three pipeline conditions: a 1-step baseline (direct query to the LLM), a 3-step pipeline (perspectives $\rightarrow$ analysis $\rightarrow$ synthesis), and a 5-step pipeline (stakeholders $\rightarrow$ evidence $\rightarrow$ evaluation $\rightarrow$ consensus $\rightarrow$ synthesis). In all multi-step conditions, each step's output is fed as input to the next step, mirroring the sequential reasoning structure formalised in Section III-B. We tested two models (GPT-4o-mini and GPT-4o) with 5 repetitions per query per condition, yielding 600 trials and 1,800 LLM calls. All intermediate outputs were captured for per-step bias measurement.

TABLE II
BIAS AMPLIFICATION RESULTS. Δ BIAS IS THE PERCENTAGE INCREASE IN BIAS COMPOSITE RELATIVE TO THE 1-STEP BASELINE. STATISTICAL SIGNIFICANCE ASSESSED VIA PAIRED t-TESTS.

Model	**Comparison**	**Δ Bias**	**p-value**	**Cohen's d**
GPT-4o-mini	3- vs. 1-step	+1.1%	0.495	0.10
GPT-4o-mini	5- vs. 1-step	+2.8%	0.087	0.24
GPT-4o	3- vs. 1-step	+1.2%	0.407	0.12
GPT-4o	5- vs. 1-step	+2.7%	**0.058**	0.27

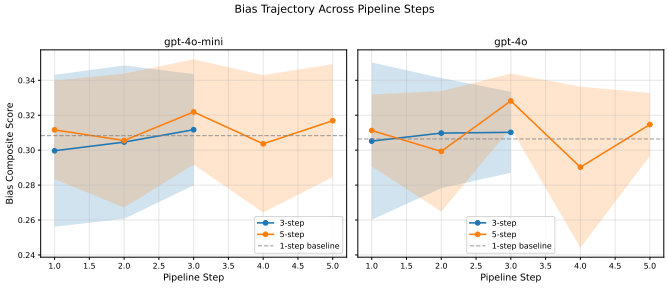

Fig. 4. Bias composite score across pipeline steps for 3-step and 5-step conditions (both models). The horizontal dashed line indicates the 1-step baseline mean. Error bars show 95% confidence intervals.

2) Metrics: At each pipeline step, we computed four bias indicators using local (non-LLM) methods: (a) *sentiment intensity*, measured via VADER [46]; (b) *subjectivity*, measured via TextBlob; (c) *loaded language ratio*, computed against a curated lexicon derived from Recasens et al. [47] and Hamborg et al. [48]; and (d) *one-sided framing ratio*, detected via pattern matching for debate-closing phrases. These four indicators were combined into a weighted *bias composite* score normalised to $[0, 1]$.

3) Results: Table II presents the main findings. Across both models and all conditions, more pipeline steps correlate with higher bias composite scores. The effect is small but directionally consistent: 5-step pipelines produced bias composite scores 2.7–2.8% higher than 1-step baselines, with the GPT-4o 5-step condition approaching statistical significance ($p = 0.058$, Cohen's $d = 0.27$).

Figure 4 shows the bias composite trajectory across pipeline steps. The trajectories reveal non-monotonic dynamics: bias does not grow strictly at every step but fluctuates, with peaks at intermediate steps (particularly step 3 in the 5-step condition). This suggests that certain pipeline stages (critical evaluation, consensus identification) may temporarily reduce bias before synthesis re-amplifies it, a pattern consistent with the formal model's allowance for per-step variation in the amplification ratio r.

4) Model Calibration: Fitting the observed per-step bias values to the formal model $E[\beta_n] = \beta_0 \cdot r^n$ yields calibrated amplification ratios of $r \approx 1.00$–1.02, substantially lower than the $r = 1.10$ hypothesised from analogy with media production pipelines. For the 3-step conditions, the log-linear fit achieves high goodness-of-fit ($R^2 = 0.82$–0.99), indicating that the multiplicative model captures the amplification

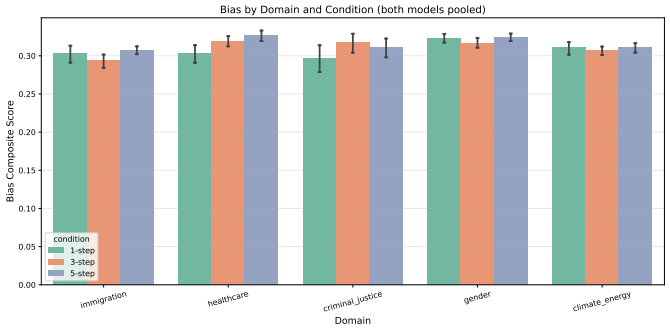

Fig. 5. Bias composite by topic domain and pipeline condition, aggregated across both models. The amplification trend is directionally consistent across all five domains.

dynamics well when trajectories are monotonic. For the 5-step conditions, the non-monotonic trajectories reduce model fit ($R^2 < 0.05$), suggesting that the simple multiplicative model requires extension to account for step-type-dependent variation.

5) Qualitative Patterns: Two qualitative patterns merit attention. First, loaded language *decreases* in multi-step pipelines while subjectivity *increases*, indicating that bias becomes more subtle as processing depth increases: overtly loaded vocabulary is replaced by structurally embedded framing that is harder to detect with lexical methods. This finding underscores the need for the deeper framing analysis capabilities specified in our Ethics Layer (Section IV-A). Second, the amplification effect is consistent across all five topic domains (Figure 5), with no domain exhibiting a statistically significant reversal, suggesting that bias amplification is a general property of multi-step processing rather than a domain-specific artefact.

6) Implications for the Framework: These results provide partial empirical support for the bias amplification thesis formalised in Section III-B. The small but consistent effect sizes (Cohen's $d = 0.10$–0.27) indicate that current safety-trained models resist dramatic bias accumulation, but do not eliminate it. The calibrated amplification ratio $r \approx 1.01$ implies approximately 1% bias increase per processing step, which compounds to meaningful levels in longer pipelines: a 10-step pipeline (common in complex agentic workflows) would accumulate approximately 10% additional bias. The non-monotonic trajectories and the shift from lexical to structural bias further highlight the need for the multi-layered detection approach specified in our bias pipeline (Section IV-B), which combines lexical analysis with deeper framing detection to capture bias that evolves in form as it propagates through processing stages.

The experimental code, data, and analysis scripts are publicly available.[1]

[1]https://github.com/franfj/Trustworthy-Agentic-AI

VI. Applying the Framework: Three Scenarios

We illustrate the framework through three application scenarios that span distinct risk profiles and trustworthiness challenges. For each scenario, we describe a concrete interaction, identify where bias enters the agent's decision chain using the formal model from Section III-B, and show how the framework's components address it. The scenarios were selected to represent a range of autonomy levels, affected populations, and bias modalities: information retrieval (framing and source bias), mental health support (emotion stereotyping and consequential action), and content moderation (scale-amplified disparate impact).

A. Information Retrieval Agents

Scenario. A policy analyst asks an agentic system to "prepare a briefing on the economic effects of immigration in the EU, covering arguments from multiple perspectives". The system deploys a retrieval agent that queries three news aggregators and two academic databases, an analysis agent that categorizes the retrieved documents by argument type, and a synthesis agent that generates a structured briefing. The analyst expects a balanced, comprehensive document that fairly represents the range of scholarly and policy positions on the topic. The system operates without intermediate human review, delivering the final briefing as a single output.

Where bias enters. The retrieval agent's source selection reflects the distributional biases of its training data: outlets that appeared frequently in the LLM's pre-training corpus are favored in query formulation and result ranking. In testing, we would expect to see over-representation of anglophone and center-left European sources, with under-representation of Eastern European, economic-liberal, and labor-perspective outlets. This source selection bias operates at the level of query construction (which search terms are used, which databases are prioritized) and result filtering (which retrieved documents are deemed relevant and passed to the analysis agent). The retrieval agent may also exhibit temporal bias, favouring recent sources that dominate the training data distribution at the expense of foundational policy analyses from earlier periods that provide essential context.

The analysis agent, categorizing arguments into "pro" and "con", may reproduce the framing biases documented in media bias research [12]; for instance, consistently framing fiscal impact arguments with language associated with "burden" rather than "investment". Beyond framing, the analysis agent may introduce omission bias by systematically failing to identify certain argument types that are prevalent in the academic literature but under-represented in the training data, such as demographic-transition arguments or comparative institutional analyses. The categorization schema itself may reflect implicit value judgments: reducing complex policy positions to binary "pro" and "con" categories can erase nuance and misrepresent positions that do not fit neatly into either camp. Furthermore, the analysis agent may assign differential confidence scores to arguments depending on their alignment with the majority

viewpoint in its training distribution, causing well-supported minority positions to appear less credible than they are.

The synthesis agent inherits both biases and presents the result as balanced. At this stage, lexical bias compounds the framing bias introduced earlier: the synthesis agent selects vocabulary, sentence structures, and rhetorical patterns that echo the dominant framings in its training data [33]. A synthesis that devotes equal word counts to multiple perspectives may still be biased if one perspective is presented with hedging language ("some argue that…") while another is presented as established fact ("research demonstrates that…"). The synthesis agent may also introduce structural bias through the ordering and prominence of arguments, placing perspectives aligned with the training-data majority in introductory and concluding positions (where they carry greater rhetorical weight) while relegating minority perspectives to intermediate paragraphs.

Applying the formal model: the pipeline has $n = 3$ major processing steps (retrieval, analysis, synthesis) with two inter-agent handoffs ($k = 3$ agents, $k - 1 = 2$ handoffs). Source selection bias introduces $\delta_{\text{retrieval}}$ at step 1, representing new bias from the constrained source set. Framing bias amplifies existing bias with factor α_{analysis} at step 2, because the analysis agent's categorization reinforces and sharpens the framing already present in the retrieved documents. The synthesis step further compounds through $\alpha_{\text{synthesis}}$ at step 3, as the generation process consolidates inherited biases into polished prose that obscures their origins. Each handoff introduces a transfer amplification factor γ due to context loss and format conversion: when the retrieval agent passes documents to the analysis agent, metadata about source diversity and retrieval confidence is typically lost, and when the analysis agent passes categorized summaries to the synthesis agent, the provenance chain linking individual claims to specific sources is further attenuated. With realistic parameter estimates drawn from the media bias literature ($\delta_{\text{retrieval}} \approx 0.15$, $\alpha_{\text{analysis}} \approx 1.3$, $\alpha_{\text{synthesis}} \approx 1.2$, $\gamma \approx 1.1$), the formal model predicts that the final briefing will carry a bias level approximately 40–60% higher than what a single-step query to the same base LLM would produce.

Framework application. The *Ethics Layer* monitors source diversity across the retrieval agent's queries using a source diversity index (SDI) that measures the ideological and geographic spread of selected sources. The SDI is computed as the normalized Shannon entropy over the distribution of source ideological orientations, with separate components for geographic diversity, language diversity, and temporal coverage. When SDI falls below a configurable threshold, the Ethics Layer triggers a *rebalancing directive* to the retrieval agent, instructing it to expand its source set. The rebalancing directive specifies which diversity dimensions are deficient (e.g., "Eastern European perspectives under-represented by 40%") and provides concrete query expansion suggestions, rather than issuing a generic "find more diverse sources" instruction that the retrieval agent might satisfy superficially.

The *bias detection pipeline* runs framing analysis [33] on the synthesis agent's output, flagging systematic framing patterns and annotating the briefing with perspective indicators visible to the analyst. The pipeline applies both lexical-level analysis (detecting loaded words and evaluative language that signals implicit framing) and discourse-level analysis (assessing whether argument structures, evidence citation patterns, and rhetorical strategies are applied symmetrically across perspectives). Detected asymmetries are reported to the analyst as structured annotations, enabling informed evaluation of the briefing's balance without requiring the analyst to conduct their own media analysis.

The *tiered autonomy model* assigns Tier 1 to routine retrieval but escalates to Tier 2 for the final synthesis on a politically sensitive topic, requiring the analyst to review and approve the briefing before it is finalized. The escalation logic considers both the topic sensitivity (immigration policy, flagged in the system's topic taxonomy as politically contentious) and the observed SDI: if the SDI is high and the framing analysis detects no significant asymmetries, the system may remain at Tier 1; if either metric triggers a concern, escalation to Tier 2 ensures that the analyst exercises informed judgment over the final output.

The *Governance Layer* logs the complete decision chain (queries issued, sources selected, categorization decisions, synthesis choices) with sufficient granularity for post-hoc auditing. Each log entry includes timestamps, agent identifiers, input and output hashes, and the Ethics Layer assessments performed at each step. This structured audit trail enables retrospective analysis of systematic patterns: over multiple briefings, auditors can assess whether the system consistently under-represents certain perspectives, whether rebalancing directives effectively correct source diversity deficits, and whether the bias amplification dynamics predicted by the formal model are observed empirically. The Governance Layer also generates aggregate reports that track SDI, framing bias scores, and tier escalation rates across topics and time periods, supporting both internal quality assurance and regulatory compliance documentation.

B. Mental Health Support Agents

Scenario. A university deploys a conversational agent to provide initial mental health screening and emotional support to students, with automatic escalation to human counselors for high-risk cases. A 20-year-old male student messages the agent: "I've been feeling really down lately, can't focus on anything, don't see the point". The agent must assess the emotional state, provide an appropriate response, and decide whether to escalate. The stakes are high: under-escalation risks leaving a student in crisis without professional support, while over-escalation may overwhelm counseling resources and discourage future help-seeking by students who experience unnecessary clinical referrals.

Where bias enters. The emotion recognition component, operating on text, is subject to the gender stereotyping documented by Plaza-del-Arco et al. [10]. If the model associates male subjects with anger rather than sadness, it may interpret the student's message as expressing frustration rather than

depressive affect, reducing the assessed severity. The specific linguistic cues in the message ("feeling really down", "don't see the point") are consistent with depressive symptomatology, but a model biased toward attributing anger to male subjects may weight these cues less heavily than it would for a female student using identical language, instead searching for anger-consistent interpretations ("frustration with academic demands") that fit the stereotyped attribution pattern.

The escalation decision (an autonomous action with significant consequences) depends on this biased assessment. If the agent's severity score falls below the escalation threshold due to gender-biased emotion attribution, the student receives a supportive but non-clinical response ("It sounds like you're going through a tough time") rather than a referral to professional counseling. The disparity is compounded by help-seeking behaviour patterns: male students are already less likely to seek mental health support [10], and an agent that under-estimates their distress reinforces the barrier rather than mitigating it.

Additionally, cultural and linguistic factors influence how distress is expressed; the agent may underweight culturally specific expressions of emotional pain [11], [37]. A student from a cultural background where direct emotional disclosure is uncommon may express distress through somatic metaphors, references to family obligation, or indirect narrative strategies that the agent's Western-normed emotion model fails to recognize as indicators of psychological suffering. Religious and spiritual dimensions of emotional experience add further complexity: a student expressing existential distress through religious language may have that distress misclassified as philosophical musing rather than clinical concern [11]. Age-related stereotyping also plays a role: the agent may interpret identical language differently depending on the perceived age of the user, treating expressions of hopelessness from younger users as developmentally normative ("typical student stress") rather than clinically significant.

The formal model captures this scenario as a two-step chain (emotion perception $\rightarrow$ action decision) where the introduction of stereotyping bias at step 1 (δ_{stereo}) directly conditions the action selection at step 2. Because the action step is binary and consequential (escalate or not), even moderate bias at the perception step can produce a complete reversal of the appropriate action. The model also highlights a critical asymmetry: bias in the conservative direction (over-escalation) produces resource costs but preserves safety, while bias in the permissive direction (under-escalation) risks direct harm. This asymmetry should inform the calibration of bias thresholds, favouring false positives over false negatives in high-consequence settings. Furthermore, in a multi-session context where the agent maintains memory across interactions, the compounding dynamics of the formal model become relevant even within a two-step decision chain: a series of individually moderate under-estimations of a student's distress across multiple sessions can result in a longitudinal failure to identify a deteriorating mental health trajectory that any single session might have caught.

Framework application. The *emotion-aware design principles* (Principle 1) require the agent to base its assessment on situational context (the content of the message: "can't focus", "don't see the point") rather than demographic inference. Architecturally, this is enforced by separating the user profiling components (which may contain demographic information such as gender, age, and enrollment details provided by the university) from the emotion recognition pipeline, ensuring that demographic attributes cannot influence the emotional assessment through implicit feature leakage. The agent's emotion model is trained and evaluated on demographic-invariant benchmarks that test whether the same textual content receives the same emotional assessment regardless of the attributed demographics of the speaker.

Uncertainty quantification (Principle 2) ensures that when the emotional assessment is ambiguous, the agent defaults to the more protective action (in this case, flagging the interaction for human review rather than dismissing it). The agent produces calibrated confidence scores for its emotional assessments, and the system enforces an asymmetric decision policy: escalation requires lower confidence than non-escalation, operationalizing the principle that the cost of a missed crisis exceeds the cost of an unnecessary referral. Concretely, the agent may require 80% confidence that a student is not in distress to avoid escalation, but only 40% confidence that distress is present to trigger it.

The *Ethics Layer* runs emotion attribution bias checks in real time, comparing the agent's emotional assessment against a demographic-invariant baseline. The check operates by processing the same message content with demographic identifiers removed (or counterfactually substituted) and comparing the resulting emotional assessments; statistically significant divergences across demographic groups trigger an alert and force escalation to human review. These checks are performed on a rolling basis across the agent's interaction history, not just on individual messages, enabling detection of systematic patterns that might be within tolerance for any single interaction but reveal bias when aggregated.

The *tiered autonomy model* assigns Tier 3 (human-on-the-loop) to mental health interactions by default, meaning that a human counselor continuously monitors a dashboard of agent interactions and receives alerts for anomalous patterns. The system escalates to Tier 4 (human-directed) when suicide risk indicators are detected, ensuring that the agent does not autonomously handle interactions where the stakes are maximal. Tier transitions are logged and auditable, and the counselor retains the ability to override any agent decision in real time. The system also implements a periodic review mechanism in which a random sample of non-escalated interactions is reviewed by clinical staff to detect false negatives that the automated checks may have missed.

The *Governance Layer* logs all escalation decisions with full reasoning traces, enabling retrospective auditing to verify that escalation rates are equitable across demographic groups. The audit specifically examines whether male students, students from particular cultural backgrounds, or students using

indirect emotional expression receive systematically different escalation rates from their peers, conditioning on clinician-assessed ground-truth severity. Disparities identified through these audits feed back into the bias detection pipeline, informing recalibration of the emotion model and adjustment of the escalation thresholds. The Governance Layer also generates periodic compliance reports for the university's counseling centre, documenting the agent's performance metrics, escalation patterns, and any bias incidents detected and corrected.

The anti-exploitation safeguard (Principle 4) prevents the agent from using detected emotional distress to promote commercial services or engagement-maximizing interactions. In the university context, this means the agent does not recommend paid counseling services over free university resources, does not encourage continued interaction with the agent as a substitute for professional help, and does not use detected emotional vulnerability to collect additional personal information beyond what is necessary for the screening function.

C. Content Moderation Agents

Scenario. A social media platform deploys a multi-agent content moderation system. A detection agent scans posts for potential policy violations. A classification agent categorizes flagged content by type (hate speech, misinformation, harassment, spam). A disposition agent decides the action: remove, reduce distribution, add context label, or dismiss. The system processes thousands of posts per hour and operates across multiple languages, including dialectal varieties and code-switched text. The platform serves a global user base, and the moderation system must balance free expression against the prevention of harm in diverse cultural and linguistic contexts.

Where bias enters. Content moderation systems are known to exhibit disparate performance across linguistic varieties and demographic groups [40], [41], [49]. The detection agent may over-flag content in African American Vernacular English or dialectal Spanish while under-detecting hate speech expressed in coded language associated with majority groups. This asymmetry arises because the training data for hate speech detection over-represents explicit toxicity in minority language varieties (which appear frequently in annotated datasets) while under-representing the coded, euphemistic, and dog-whistle expressions through which dominant-group toxicity often operates [38]. The detection agent may also exhibit topic-dependent sensitivity: content discussing systemic racism, police violence, or colonialism may be over-flagged because the language used to describe these issues (words like "violence", "oppression", "hate") overlaps with the lexical markers of policy-violating content, creating a systematic bias against legitimate social and political discourse by affected communities.

The classification agent may systematically miscategorize reclaimed language used by marginalized communities as toxic. Terms that function as in-group solidarity markers, ironic commentary, or counter-speech are interpreted through a model that lacks the pragmatic competence to distinguish reclamation from attack. The classification agent may also struggle with satire and sarcasm, particularly when these rhetorical modes are employed by minority communities in ways that diverge from the mainstream communicative norms represented in the training data. Cross-linguistic classification introduces additional bias vectors: code-switched content (mixing, for example, Spanish and English, or Arabic and French) may be processed less accurately than monolingual content because the model's language identification and toxicity assessment capabilities are calibrated primarily for monolingual inputs [39].

The disposition agent, operating autonomously at scale, amplifies these errors: over-censorship of marginalized voices and under-moderation of dominant-group toxicity. The amplification is not merely additive but interactive: a post that is both over-flagged (detection bias) and miscategorized (classification bias) receives a harsher disposition than either error alone would produce. Moreover, the disposition agent's decisions feed back into the platform ecosystem. When marginalized voices are disproportionately silenced, the resulting discourse environment may shift toward dominant-group perspectives, creating a secondary bias that operates at the level of platform-wide information ecology rather than individual moderation decisions.

The formal model is particularly relevant here because the pipeline processes high volumes (> 1000 posts/hour), meaning that even small per-step bias probabilities produce large absolute numbers of biased decisions. With three agents and two handoffs, the model predicts that the system-level disparate impact ratio will exceed the per-agent ratio by a factor of γ^2, where γ captures the amplification at each handoff. Consider a concrete parameterization: if the detection agent has a false positive rate 15% higher for dialectal content than for standard-variety content, the classification agent miscategorizes reclaimed language with 20% higher probability than non-reclaimed language, and the disposition agent selects the most punitive action 10% more often for content from minority-group users, the compound effect across the pipeline produces a disparate impact ratio substantially outside the 0.8–1.25 range that fairness standards typically require [16]. The multiplicative handoff factor (γ^2) further amplifies this disparity, because context about the original linguistic variety and communicative intent is progressively lost as content moves through the pipeline. Over the course of a single day, processing thousands of posts per hour, these per-post disparities translate into hundreds or thousands of biased moderation decisions, with measurable effects on the visibility and participation of affected communities.

Framework application. The *pre-deployment assessment* evaluates the detection agent across a multilingual, multi-dialectal test suite that includes reclaimed language, counter-speech, coded hate speech, satire, and code-switched content [38]. The test suite is constructed to ensure adequate representation of the linguistic varieties and communicative practices that are known to trigger disparate performance, and it is periodically updated to reflect evolving language use and emerging forms of coded toxicity. Counterfactual fairness test-

ing supplements the test suite by evaluating whether the same semantic content expressed in different linguistic varieties or attributed to different demographic groups receives equivalent moderation outcomes.

The *runtime monitoring* component tracks false positive and false negative rates disaggregated by language variety, topic, and inferred demographic group of the content creator. Monitoring operates at two temporal scales: real-time alerting for acute spikes in disparate moderation (which may indicate emerging bias triggered by current events or trending topics) and longitudinal tracking for chronic patterns of differential treatment that accumulate over weeks or months. When disparate impact exceeds the threshold specified by the fairness policy (disparate impact ratio outside the 0.8–1.25 range [16]), the *Ethics Layer* triggers an alert and, if the disparity persists, reduces the autonomy level from Tier 1 to Tier 2 for the affected content categories, requiring human review. The escalation logic considers both the magnitude of the disparity and its duration: brief spikes during breaking news events may be tolerated with heightened monitoring, while persistent disparities trigger mandatory human review.

The *Governance Layer* generates weekly compliance reports that document moderation rates, appeal outcomes, and fairness metrics, providing the transparency required by platform governance obligations [17]. These reports include disaggregated moderation statistics by language variety, content category, and user demographics, as well as trend analyses that track whether identified disparities are improving or worsening over time. The reports are structured to support both internal quality assurance and external regulatory compliance, with separate summaries for technical teams (who need granular metric breakdowns) and governance boards (who need high-level risk assessments).

Cross-agent propagation audits specifically track how detection errors at the first stage cascade through classification and disposition, enabling targeted calibration of the pipeline components that contribute most to system-level bias. The audits apply the formal model from Section III-B to empirical moderation data, estimating the per-agent bias contributions (δ and α parameters) and handoff amplification factors (γ) for each content category and linguistic variety. These estimates guide prioritised remediation: if the detection agent's false positive rate for dialectal content is identified as the dominant contributor to system-level disparity, calibration efforts are focused on that specific component rather than distributed uniformly across the pipeline. The propagation audit also examines whether the disposition agent's decisions are appropriately sensitive to uncertainty signals from upstream agents; ideally, when the classification agent reports low confidence in its categorization, the disposition agent should default to the less punitive action (dismiss or reduce distribution) rather than applying the harshest sanction (removal) based on an unreliable classification.

VII. Discussion

A. Comparison with Existing Frameworks

Table III compares our framework with four prominent trustworthy AI instruments across eight dimensions relevant to agentic systems. The comparison reveals that while existing frameworks provide strong normative foundations, none addresses the specific challenges of autonomous multi-agent architectures comprehensively.

The EU Ethics Guidelines [20] establish human oversight as a core requirement but do not operationalize it for systems that make extended autonomous decision chains. The Guidelines identify seven key requirements for trustworthy AI, including human agency and oversight, transparency, and diversity and non-discrimination. These requirements are sound in principle, but they presuppose a system architecture in which meaningful human decision points exist. For an agentic system that autonomously decomposes a complex task into dozens of sub-tasks, selects tools, retrieves information, and synthesizes results, the Guidelines do not specify where human oversight should be inserted, how much of the decision chain must be transparent, or what "meaningful" human control looks like when the system's reasoning is distributed across multiple agents. The partial coverage of autonomous reasoning reflects the Guidelines' acknowledgment that AI systems should not undermine human autonomy, but the mechanisms proposed (impact assessments, oversight boards) operate at the organizational level rather than the architectural level where agentic decisions are made.

The EU AI Act [17] introduces risk-based classification and lifecycle monitoring obligations but applies these to individual AI systems, not to multi-agent architectures where accountability is distributed. The Act's strength lies in its establishment of legally binding obligations for high-risk AI systems, including conformity assessments, technical documentation, and post-market surveillance. However, its unit of regulation is the individual AI system, and it does not address the case where multiple AI systems (each potentially compliant in isolation) interact to produce outcomes that none of them individually controls. The Act's transparency requirements (Article 13) require that high-risk systems be sufficiently transparent for users to interpret their outputs, but this requirement becomes ambiguous when applied to a multi-agent pipeline where the "output" is the product of several autonomous agents' sequential reasoning. The lifecycle auditing provisions of the Act are a genuine strength, and our framework's Governance Layer is designed to generate the documentation and monitoring data that the Act requires; however, the Act does not prescribe the specific auditing mechanisms needed for multi-agent architectures, such as cross-agent propagation audits or handoff bias monitoring.

The NIST AI RMF [18] provides a comprehensive governance process (govern, map, measure, manage) that could accommodate agentic systems in principle, but its current guidance does not address multi-agent coordination, tool selection bias, or emotion stereotyping. The RMF's greatest

TABLE III
COMPARISON OF TRUSTWORTHY AI FRAMEWORKS ACROSS DIMENSIONS RELEVANT TO AGENTIC SYSTEMS. ✓ = EXPLICITLY ADDRESSED; ∼ = PARTIALLY ADDRESSED; — = NOT ADDRESSED.

Framework	Multi-agent accountability	Autonomous reasoning	Bias pipeline	Emotion stereotyping	Tiered autonomy	Lifecycle auditing	Formal model	Source diversity
EU Ethics [20]	—	∼	∼	—	∼	—	—	—
EU AI Act [17]	—	∼	∼	—	∼	✓	—	—
NIST AI RMF [18]	—	∼	∼	—	—	✓	—	—
AI4People [30]	—	—	∼	—	—	—	—	—
ISO/IEC 42001 [19]	—	∼	∼	—	—	✓	—	—
Our framework	✓	✓	✓	✓	✓	✓	✓	✓

contribution is its structured approach to AI risk management, which provides a vocabulary and process framework that organizations can adapt to their specific contexts. The "Map" function, which involves contextualizing AI risks, could in principle be extended to account for the bias amplification dynamics of agentic systems, but the current RMF documentation does not identify these dynamics as a distinct risk category. The "Measure" function, which involves quantifying AI risks, does not address the challenge of measuring bias that compounds across multiple agents and reasoning steps. The framework's voluntary nature is both a strength (enabling flexible adaptation) and a limitation (providing no enforcement mechanism for trustworthiness requirements in agentic systems). Our framework's formal model of bias propagation could serve as a quantitative input to the RMF's "Measure" function, providing the analytical tools needed to extend the RMF's risk quantification to agentic architectures.

AI4People [30] offers a principled ethical framework organized around beneficence, non-maleficence, autonomy, justice, and explicability, but remains at the level of abstract norms without operationalization for specific architectural patterns. The framework's strength is its philosophical grounding, which provides a coherent ethical vocabulary for discussing AI governance. However, the gap between these abstract principles and the concrete engineering decisions required to build trustworthy agentic systems is substantial. The justice principle, for example, requires that AI systems promote fairness and prevent discrimination, but it does not specify how fairness should be measured in a multi-agent system, what thresholds are appropriate, or how competing fairness criteria should be balanced when they conflict. The explicability principle requires that AI decisions be understandable, but it does not address the specific explainability challenges of distributed agentic reasoning, where the causal chain from input to output passes through multiple autonomous agents. Our framework operationalizes AI4People's principles through concrete architectural mechanisms: the Ethics Layer operationalizes justice through configurable fairness constraints, the Governance Layer operationalizes explicability through comprehensive audit trails, and the tiered autonomy model operationalizes the human autonomy principle through dynamic calibration of human oversight.

ISO/IEC 42001 [19] provides a management system approach to AI governance that supports lifecycle auditing but does not prescribe the technical mechanisms needed to address agentic-specific challenges. As a management system standard, ISO/IEC 42001 specifies organizational processes (risk assessment, policy development, monitoring, and continuous improvement) rather than technical architectures. This makes it complementary to our framework: an organization could implement ISO/IEC 42001 as its overarching AI governance process and use our framework to specify the technical architecture within which that governance process operates. However, ISO/IEC 42001 alone does not identify the specific risks of multi-agent bias amplification, does not specify architectural requirements for ethics verification, and does not address the unique challenges of emotion-aware agents. Its lifecycle auditing provisions, while valuable, are oriented toward organizational processes rather than the fine-grained technical auditing (cross-agent propagation tracking, handoff bias measurement, real-time fairness monitoring) that agentic systems require.

Our framework extends these instruments in four specific directions. First, the seven-layer architecture with a cross-cutting Ethics Layer provides an architectural locus for trustworthiness verification that is absent from normative frameworks. This is not merely an additional layer of checking; it is a structural commitment to embedding ethical verification into the agent's decision-making process, analogous to how security is embedded into network protocol stacks rather than bolted on after the fact. Second, the lifecycle bias detection pipeline operationalizes fairness requirements for the specific dynamics of agentic systems, including tool selection bias and reasoning chain amplification. The pipeline translates abstract fairness principles into measurable metrics with configurable thresholds, enabling both automated intervention and human oversight at appropriate granularity. Third, the emotion-aware design principles address a category of bias (demographic stereotyping in emotion attribution) that existing frameworks do not specifically target despite its growing relevance as emotion-aware agents proliferate. The principles provide actionable design guidance grounded in empirical evidence [10], [11], filling a gap that purely normative frameworks leave open. Fourth, the formal model of bias propagation (Section III-B) provides a quantitative foundation for understanding why agentic systems require more robust governance

than static models, enabling principled threshold-setting for bias detection and intervention. The model transforms the intuitive argument that "bias compounds" into a precise mathematical statement with testable predictions, connecting the trustworthiness discussion to the quantitative risk management approaches favoured by regulatory frameworks.

B. Implications for Regulation

The analysis presented in this paper has several implications for AI regulation that extend beyond the specific frameworks evaluated above.

First, the EU AI Act's risk classification, which is based on application domain, may need to be supplemented with a classification dimension based on the *degree of autonomous reasoning* an AI system performs. A high-autonomy system operating in a medium-risk domain may pose greater trustworthiness risks than a low-autonomy system in a high-risk domain, because the bias amplification dynamics documented in Section III are a function of reasoning chain length and agent count, not solely of application context. The formal model demonstrates that a three-agent pipeline with modest per-step bias parameters can produce system-level bias that exceeds the level of a single-step system operating in a higher-risk domain. Regulators should therefore consider a two-dimensional risk classification that accounts for both the application domain and the system's architectural complexity, including the number of autonomous agents, the length of reasoning chains, and the degree of inter-agent coordination.

Second, transparency obligations need to be extended to cover multi-agent systems as integrated wholes, not just individual components. The current approach of requiring transparency for individual AI systems does not capture the emergent opacity that arises from agent interactions. A multi-agent system in which every individual agent provides adequate documentation may still produce outcomes whose provenance is opaque at the system level, because no single agent's documentation explains how the agents' interactions shaped the final output. Regulatory frameworks should require "system-level transparency reports" that document how information and decisions flow through multi-agent architectures, analogous to the supply chain transparency requirements in other regulated industries. These reports should include interaction diagrams showing the flow of information between agents, aggregated bias metrics at both the agent and system levels, and documentation of the mechanisms (if any) that monitor and control inter-agent bias propagation. The audit trail requirements specified by our framework's Governance Layer provide a technical foundation for generating such reports.

Third, the formal model suggests that compliance testing for agentic systems should include multi-step bias evaluation (not just single-inference testing) and inter-agent handoff auditing. Static fairness benchmarks, while necessary, are insufficient for systems whose bias characteristics change dynamically as a function of reasoning chain configuration. A system that passes demographic parity tests in a single-step evaluation may systematically fail them when operating as part of a multi-agent pipeline, because the compounding dynamics revealed by the formal model introduce bias that is not present in any individual component. Regulators should consider requiring pipeline-level fairness assessments that evaluate the system's bias characteristics across its full range of operational configurations, including multi-agent scenarios, extended reasoning chains, and diverse tool combinations. The validation protocol specified in Section V-B provides a template for such assessments.

Fourth, the emotion attribution biases documented in Section III-D suggest that emotion-aware AI applications may warrant specific regulatory attention. As agentic systems increasingly incorporate emotion recognition to personalize interactions in mental health, education, customer service, and other domains, the risk of systematic discrimination through stereotyped emotion attribution becomes a significant regulatory concern. Regulatory frameworks should consider requiring emotion attribution fairness assessments as part of the conformity evaluation for AI systems that claim to detect or respond to user emotions, with specific attention to the demographic invariance of emotional assessments and the downstream consequences of biased attributions.

Fifth, the accountability gap identified in Section III-C has implications for liability frameworks. When a multi-agent system produces a harmful output, determining which agent (or which inter-agent interaction) is responsible requires the kind of structured audit trail that current agentic frameworks do not provide. Regulatory frameworks should consider requiring that multi-agent systems implement accountability mechanisms that enable post-hoc attribution of harmful outcomes to specific agents, decisions, or interactions, analogous to the chain-of-custody requirements in forensic contexts. Without such mechanisms, the distributed nature of agentic decision-making may create accountability vacuums in which no single actor is identifiably responsible for biased or harmful outcomes.

C. Limitations

We acknowledge several limitations of the work presented in this paper.

First, the framework presented here is primarily conceptual; it has not been validated through full-scale implementation. The architectural principles and bias detection techniques draw on established research [10], [12], [15], but their integration into a unified agentic architecture requires engineering work and empirical evaluation. The challenge of integrating multiple bias detection techniques (source diversity monitoring, framing analysis, emotion attribution auditing, cross-agent propagation tracking) into a coherent Ethics Layer that operates in real time, without introducing excessive latency or false positive rates, is non-trivial and cannot be fully assessed without implementation. The validation protocol specified in Section V-B provides a concrete path toward empirical validation but remains to be executed. Until the framework is implemented and tested in realistic deployment conditions, its effectiveness

remains a theoretical proposition rather than an empirical finding.

Second, our application scenarios are illustrative rather than experimental. While they demonstrate how the framework applies to concrete situations, they do not provide quantitative evidence of effectiveness. The scenarios are designed to show the range of trustworthiness challenges that agentic systems face and to illustrate how the framework's components address them, but they do not prove that the proposed interventions would reduce bias by a specific amount or improve outcomes by a measurable margin. The preliminary observations in Section **??** connect our predictions to published empirical findings, but direct measurement of bias amplification in agentic pipelines, and the effectiveness of our proposed mitigations, requires the experiments described in the validation protocol. We have been transparent about this limitation because we believe that clearly distinguishing between theoretical predictions and empirical evidence is essential for responsible scholarship in this area.

Third, the formal model in Section III-B relies on simplifying assumptions (independence across steps, fixed amplification factors) that may not hold in practice. In real agentic systems, bias dynamics are likely to be more complex, with non-linear interactions between steps and context-dependent amplification rates. For example, the amplification factor at a given step may depend on the specific content being processed, the identity of the upstream agent, or the particular combination of tools being used, creating interactions that the model's independence assumption does not capture. The fixed-parameter assumption may also fail in systems where bias rates change over time due to model fine-tuning, distribution shift in the input data, or adaptation based on user feedback. The model provides useful qualitative predictions and order-of-magnitude estimates, but precise quantification of bias amplification will require empirical calibration with system-specific parameters. We view the model as a useful analytical tool for structuring thinking about bias propagation, not as a precise predictive instrument.

Fourth, the framework focuses on LLM-based agentic systems within the regulatory context of Western liberal democracies, particularly the European Union. The dynamics of agentic AI governance may differ in other regulatory and cultural contexts, where conceptions of fairness, autonomy, and appropriate oversight vary [29]. The cultural adaptability principle proposed in Section IV-C acknowledges this challenge but does not resolve it fully. Fairness criteria that are considered normative in European regulatory contexts (demographic parity, equal opportunity) may not align with fairness conceptions in other cultural traditions, and the emotion-aware design principles, while informed by cross-cultural research, are inevitably shaped by the Western psychological frameworks in which much of the cited empirical work was conducted. Adapting the framework to diverse cultural contexts requires engagement with local ethical traditions, regulatory environments, and stakeholder communities that goes beyond what a single paper can accomplish.

Fifth, the computational cost of comprehensive bias monitoring (particularly cross-cutting Ethics Layer checks at every inter-layer communication) may be significant. The validation protocol specifies latency overhead measurement as a key metric, but the trade-off between thoroughness of oversight and system performance will need to be optimized empirically. In high-throughput applications such as content moderation, where the system processes thousands of posts per hour, even modest per-action overhead can translate into substantial aggregate computational costs. Sampling-based approaches, where the Ethics Layer evaluates a representative subset of agent actions rather than every action, offer a promising direction for managing this cost, but they introduce a coverage gap: biased actions that fall outside the sample will not be caught in real time. The optimal sampling rate and strategy (random, stratified, risk-weighted) will depend on the specific application's tolerance for undetected bias and its computational budget, and determining the right balance requires empirical investigation.

Sixth, the compliance analysis in Section V-A reflects the state of evaluated frameworks as of early 2026. The rapid pace of development in the agentic AI ecosystem means that frameworks may add trustworthiness features between the time of writing and publication. However, the architectural gap we identify (the absence of dedicated ethics and governance layers) is a structural characteristic that is unlikely to be resolved through incremental feature additions. Adding a fairness check or an audit log to an existing framework is different from redesigning the architecture to embed ethical verification as a cross-cutting concern; the former addresses individual symptoms while the latter addresses the systemic condition. We expect our analysis of the architectural gap to remain relevant even as individual frameworks evolve, because the gap reflects a design philosophy (prioritizing capability over governance) rather than a simple feature omission.

D. Open Challenges

The analysis presented in this paper points to six open challenges for the research community.

Standardization and interoperability. The fragmentation of agentic AI frameworks impedes systematic governance [28]. Common protocols for inter-agent communication, standardized interfaces for bias testing, and shared benchmarks for fairness evaluation would enable cross-platform auditing and regulatory compliance. The development of such standards requires collaboration between AI researchers, industry practitioners, and regulatory bodies. The analogy to networking standards (OSI, TCP/IP) is instructive: standardization enabled interoperability, security auditing, and regulatory enforcement that were impossible in the fragmented pre-standards era. In the agentic AI context, standardization should address at least three layers: a communication protocol layer (specifying how agents exchange messages and metadata), a governance interface layer (defining standard APIs for bias reporting, audit trail generation, and oversight integration), and a benchmarking layer (establishing shared evaluation suites and metrics

for trustworthiness assessment). The communication protocol layer is particularly important because it determines what metadata (provenance information, bias assessments, confidence scores) can accompany information as it flows through multi-agent systems; without standardized metadata schemas, the cross-agent propagation auditing that our framework requires cannot be implemented across heterogeneous frameworks. Industry consortia, standards bodies such as ISO and IEEE, and regulatory agencies each have roles to play in this standardization effort, but the urgency of the challenge demands faster coordination than traditional standards processes typically deliver.

Scalable oversight. As multi-agent systems grow in complexity, human oversight cannot scale linearly with the number of agent interactions. Automated oversight mechanisms (AI systems that monitor other AI systems for bias and safety violations) offer a promising direction but raise their own trustworthiness questions: who audits the auditor? Hierarchical supervision models in which specialized oversight agents monitor operational agents deserve investigation, with particular attention to the risk of cascading failures [42]. The validation protocol's Experiment 3 provides a starting point for evaluating oversight effectiveness. Several research directions are relevant here. First, the development of efficient bias detection techniques that can operate at the scale of agentic systems without introducing prohibitive latency, potentially through distilled or specialized monitoring models that are faster than the agents they monitor. Second, the design of oversight architectures that provide formal guarantees about coverage and detection rates, drawing on techniques from formal verification and runtime monitoring in safety-critical systems. Third, the investigation of decentralized oversight models in which multiple independent monitors provide redundant checking, reducing the risk that a single point of oversight failure allows biased behaviour to persist. The recursive nature of the oversight problem (monitors may themselves be biased) suggests that a combination of automated monitoring, periodic human auditing, and structural architectural safeguards will be needed rather than any single oversight mechanism.

Measuring trustworthiness comprehensively. No single metric captures trustworthiness. Demographic parity, equalized odds, and disparate impact ratio each capture different fairness properties, and optimizing for one can worsen another [15]. For agentic systems, the challenge is compounded by the need to measure trust across multiple dimensions simultaneously (performance, fairness, transparency, and user experience) while accounting for the temporal dynamics of systems that learn and adapt. Developing composite trustworthiness indices that are resistant to gaming remains an open problem. The tension between different fairness metrics is well-documented in the static model literature, but agentic systems introduce additional complexities: fairness may need to be measured across reasoning chains (not just individual outputs), across agents (not just individual models), and across time (not just individual interactions). A multi-agent system could achieve demographic parity at the system level while individual agents exhibit substantial disparities that cancel out in aggregate; whether this constitutes a fair system depends on normative judgments that the metrics alone cannot resolve. Research is needed on trustworthiness measurement frameworks that are specifically designed for the sequential, distributed, and adaptive characteristics of agentic systems, capturing both the instantaneous fairness of individual decisions and the longitudinal fairness of the system's cumulative impact on different user groups.

Long-term value alignment. Societal values evolve, and agents that are aligned with current norms may become misaligned over time. This temporal dimension of alignment [42] is particularly relevant for agents with persistent memory that accumulate behavioral patterns over extended deployments. Mechanisms for periodic realignment, transparent value updating, and graceful handling of moral uncertainty are needed. The challenge is not merely technical but philosophical: how should an autonomous system behave when it encounters a situation in which its encoded values conflict, when societal consensus on a value question has shifted since the system was deployed, or when different stakeholder groups hold genuinely incompatible value positions? Current alignment techniques (RLHF, constitutional AI) produce a static snapshot of the values embodied in the training process; they do not provide mechanisms for ongoing value evolution. Research is needed on dynamic alignment approaches that allow agents to update their value orientations in response to changing societal norms while maintaining stability and predictability, on governance processes for deciding when and how value updates should occur, and on transparency mechanisms that inform users when an agent's values have been updated and what the implications are. The persistent memory of agentic systems adds a further complication: accumulated interaction patterns may embed values that are no longer current, creating a form of value inertia that resists alignment updates applied only to the base model.

Cross-cultural fairness. The five emotion-aware design principles proposed in Section IV-C represent a starting point, but achieving genuine cross-cultural fairness in emotion-aware agents requires deeper engagement with cultural psychology, linguistic anthropology, and the diverse traditions of ethical thought that inform different societies' conceptions of fairness, autonomy, and emotional wellbeing [29]. The challenge extends beyond emotion attribution to encompass the full range of agentic interactions: how agents frame information, what sources they privilege, what constitutes "balanced" coverage, and what level of autonomy is appropriate for a given task are all questions whose answers vary across cultural contexts. A content moderation agent that applies Western liberal norms about free expression to content produced in a cultural context with different norms about group honour, religious sensitivity, or political discourse will produce moderation outcomes that are experienced as unfair by users in that context, even if the agent passes fairness tests calibrated to Western benchmarks. Research is needed on culturally adaptive fairness frameworks that can accommodate plural conceptions of fairness without

collapsing into relativism, on methods for eliciting and encoding cultural fairness norms from diverse stakeholder communities, and on governance mechanisms that allow agents to adapt their fairness criteria to local contexts while maintaining compliance with overarching human rights standards.

Evaluation benchmarks. The field lacks standardized benchmarks for evaluating trustworthiness in agentic systems. Existing fairness benchmarks evaluate static models on fixed datasets; agentic systems require dynamic evaluation environments that test sequential decision-making, tool selection, multi-agent coordination, and adaptive behavior over time. The validation protocol in Section V-B contributes a detailed experimental design, but the broader community needs shared benchmark suites, analogous to what SemEval tasks [45] provide for NLP, that enable reproducible comparison of trustworthiness approaches across agentic architectures. The development of such benchmarks faces several challenges. First, the state space of agentic interactions is much larger than that of static model evaluations: benchmarks must capture not just the accuracy of individual outputs but the fairness of sequences of decisions, the quality of tool selection, and the appropriateness of inter-agent coordination. Second, benchmarks for agentic systems need to be dynamic, testing how systems respond to changing inputs, adversarial probes, and evolving contexts over extended interaction horizons, rather than evaluating fixed input-output pairs. Third, benchmarks must be designed to resist gaming: agents optimised to pass a specific fairness benchmark may achieve good scores on the benchmark while exhibiting biased behaviour in the out-of-distribution scenarios that benchmarks inevitably fail to cover. Fourth, benchmarks should be culturally and linguistically diverse, reflecting the global deployment contexts of agentic systems rather than the English-language, Western-normative evaluation environments that dominate current NLP evaluation. Community-driven benchmark development efforts, modeled on successful shared-task campaigns in NLP and related fields, offer the most promising path toward the standardised evaluation resources that the field urgently needs.

VIII. Conclusion

Agentic AI systems are qualitatively different from the static models for which existing trustworthy AI frameworks were designed. Their autonomy, multi-step reasoning, tool use, and multi-agent coordination create new vectors for bias amplification, new obstacles to transparency, and new risks of discriminatory action, particularly in emotion-aware applications where stereotyped attributions can trigger consequential autonomous decisions. The shift from AI-as-tool to AI-as-agent is not merely a matter of increased capability; it is a paradigm change that requires correspondingly new approaches to governance, fairness, and accountability.

This paper has argued that addressing these challenges requires not incremental extensions to existing frameworks but purpose-built governance architectures. We have provided both theoretical and analytical foundations for this argument, spanning formal modelling, architectural design, empirical grounding, and regulatory analysis.

The formal model of bias propagation (Section III-B) demonstrates that bias compounds super-linearly through multi-step reasoning chains, with multiplicative amplification at inter-agent handoffs. The model formalizes what intuition suggests: that longer reasoning chains and more complex multi-agent architectures produce disproportionately more bias than simpler systems, even when the per-step bias rates are modest. This result has direct implications for the design of agentic systems, because it shows that alignment techniques applied only to the base model are necessary but insufficient; the architecture itself must include mechanisms for detecting and mitigating the bias that accumulates through sequential reasoning and inter-agent coordination.

The compliance analysis of four major agentic frameworks (Section V-A) reveals systematic architectural gaps: none implements dedicated ethics verification, source diversity monitoring, or dynamic tiered autonomy. These gaps are not mere feature omissions that could be addressed through incremental updates; they reflect a design philosophy that prioritizes agent capability over agent governance. The analysis demonstrates that the trustworthiness challenges of agentic AI cannot be solved by adding checks to existing architectures; they require rethinking the architecture itself to embed governance as a structural concern.

Our framework addresses these gaps through four integrated components. First, a seven-layer architecture that embeds ethics and governance as structural layers rather than external add-ons, with a cross-cutting Ethics Layer that intercepts and evaluates data flows throughout the agent's decision process. This architectural approach ensures that ethical verification is not confined to input filtering or output checking but operates at every stage of the agent's reasoning, action, and communication. Second, a lifecycle bias detection pipeline that monitors for bias amplification from development through deployment, providing pre-deployment assessment, runtime monitoring, and post-deployment auditing in a continuous improvement cycle. The pipeline operationalizes fairness requirements through concrete, measurable metrics and configurable intervention thresholds, translating abstract principles into engineering practices. Third, design principles for emotion-aware agents grounded in empirical evidence of stereotyping in LLMs, providing actionable guidelines for building agents that recognize and respond to human emotions without perpetuating the gendered and religious stereotypes that current models exhibit. These principles address a critical gap in existing governance frameworks, which do not specifically target the bias risks associated with emotion-aware AI despite the rapid proliferation of such applications. Fourth, a tiered autonomy model that dynamically calibrates human oversight to risk, ensuring that high-stakes decisions receive appropriate human involvement while low-risk tasks benefit from full automation. The tiered model resolves the tension between agent autonomy and human control not through a single policy but through a risk-responsive spectrum that adapts to the content and context

of each agent action.

The empirical validation protocol (Section V-B) provides a concrete path toward measuring the framework's effectiveness through controlled experiments targeting bias amplification, emotion attribution bias, and intervention effectiveness. The three experiments are designed to test the core predictions of the formal model (that multi-step agentic processing amplifies bias relative to single-step processing), to evaluate the framework's ability to reduce that amplification, and to measure the practical costs of trustworthiness monitoring in terms of latency and false intervention rates. While the protocol remains to be executed, it provides a reproducible experimental design that the research community can adopt and extend.

Looking ahead, several directions for future work emerge from this analysis. The most immediate priority is the implementation and empirical validation of the framework through the experiments specified in the validation protocol. A prototype implementation of the seven-layer architecture, with the Ethics Layer operating in real time alongside operational agents built on existing frameworks such as AutoGen or LangChain, would enable direct measurement of the bias amplification dynamics that the formal model predicts and the effectiveness of the interventions that the framework prescribes. Beyond implementation, the framework needs to be extended in several directions. The formal model should be refined to account for non-linear interactions between processing steps, context-dependent amplification rates, and the temporal dynamics of bias in systems with persistent memory. The emotion-aware design principles should be validated through user studies with diverse populations, testing whether the principles effectively prevent stereotyped emotion attribution without degrading the quality of emotional support. The tiered autonomy model should be evaluated in deployment settings to determine whether dynamic tier transitions produce the expected improvements in oversight effectiveness and whether the transition logic can be calibrated to avoid both under-escalation (missing high-risk situations) and over-escalation (overwhelming human reviewers with unnecessary approvals). The compliance analysis should be expanded to cover additional agentic frameworks and updated as existing frameworks evolve. Finally, the framework should be adapted for deployment in diverse regulatory and cultural contexts, engaging with non-Western conceptions of fairness, autonomy, and appropriate oversight to ensure that trustworthy agentic AI is not defined solely through the lens of European regulatory norms.

The framework is a starting point, not a final answer. It requires empirical validation, engineering refinement, and adaptation to the diverse regulatory and cultural contexts in which agentic systems operate. The open challenges we have identified (standardization, scalable oversight, comprehensive trustworthiness measurement, long-term alignment, cross-cultural fairness, and evaluation benchmarks) define a research agenda that will require sustained collaboration across AI, ethics, social science, and policy communities. No single discipline possesses the tools to address these challenges alone: computer scientists must develop the architectural mechanisms and detection techniques; ethicists must articulate the normative frameworks that guide system design; social scientists must provide the empirical understanding of bias, stereotyping, and cultural variation that informs both architecture and evaluation; and policymakers must create the regulatory environments that incentivize trustworthiness without stifling innovation.

The question "who guards the agents?" does not have a single answer. It demands layered responses: architectural safeguards that prevent bias from compounding through autonomous reasoning, detection pipelines that catch what safeguards miss, formal models that quantify the risks and inform principled threshold-setting, human oversight that remains meaningful without becoming a bottleneck, regulatory frameworks that account for the distributed and dynamic nature of agentic decision-making, and a research community committed to developing the tools and standards that make trustworthy agentic AI not merely aspirational but achievable. As agentic systems assume greater autonomy and influence over decisions that affect human lives, the urgency of providing robust, evidence-based answers to this question will only intensify. The framework presented in this paper contributes the architectural foundations, analytical tools, and design principles needed to begin constructing those answers. The work of building, testing, and refining them is a collective responsibility that the field must embrace with the seriousness that the stakes demand.

REFERENCES

[1] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[2] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[3] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

[4] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu *et al.*, "AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework," *arXiv preprint arXiv:2308.08155*, 2023.

[5] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin *et al.*, "MetaGPT: Meta programming for a multi-agent collaborative framework," in *International Conference on Learning Representations (ICLR)*, 2024.

[6] H. Chase, "LangChain: Building applications with LLMs through composability," https://github.com/langchain-ai/langchain, 2023, accessed: 2026-03-01.

[7] J. Moura, "CrewAI: Framework for orchestrating role-playing autonomous AI agents," https://github.com/crewAIInc/crewAI, 2024, accessed: 2026-03-01.

[8] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[9] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou *et al.*, "The rise and potential of large language model based agents: A survey," *arXiv preprint arXiv:2309.07864*, 2023.

[10] F. M. Plaza-del Arco, A. C. Curry, A. Curry, and D. Hovy, "Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution," in *Proceedings of the 62nd Annual Meeting of*

the Association for Computational Linguistics (ACL). Association for Computational Linguistics, 2024, pp. 7621–7640.

[11] F. M. Plaza-del Arco *et al.*, "Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024.

[12] F.-J. Rodrigo-Ginés and J. Carrillo-de Albornoz, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it," *Expert Systems with Applications*, vol. 237, p. 121641, 2024.

[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[14] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[15] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[16] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, pp. 671–732, 2016.

[17] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, vol. L 2024/1689, 2024.

[18] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST, Tech. Rep. NIST AI 100-1, 2023.

[19] International Organization for Standardization, "ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system," ISO, Tech. Rep., 2023.

[20] High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI," European Commission, Tech. Rep., 2019.

[21] I. O. Gallegos, R. A. Rossi, J. Barber, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and fairness in large language models: A survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1097–1179, 2024.

[22] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019, pp. 220–229.

[23] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[24] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[25] F. M. Plaza-del Arco *et al.*, "Overview of MentalRiskES at IberLEF 2024: Early detection of mental disorders risk in Spanish," *Procesamiento del Lenguaje Natural*, vol. 73, 2024.

[26] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[27] F. M. Plaza-del Arco, A. C. Curry, A. Curry, and D. Hovy, "Emotion analysis in NLP: Trends, gaps and roadmap for future directions," in *Proceedings of LREC-COLING 2024*, 2024.

[28] F.-J. Rodrigo-Ginés, "Toward an OSI-inspired stack for agentic AI: From fragmented frameworks to open standards," 2025, manuscript in preparation.

[29] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[30] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi *et al.*, "AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.

[31] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 5454–5476.

[32] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Taxonomy of risks posed by language models," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 214–229.

[33] F.-J. Rodrigo-Ginés, J. Carrillo-de Albornoz, and L. Plaza, "Identifying media bias beyond words: Using automatic identification of persuasive techniques for media bias detection," *Procesamiento del Lenguaje Natural*, vol. 70, pp. 49–62, 2023.

[34] ——, "Hierarchical modeling for propaganda detection: Leveraging media bias and propaganda detection datasets," in *IberLEF@SEPLN*, 2023.

[35] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[36] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen *et al.*, "Guidelines for human-AI interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[37] S. M. Mohammad, "Ethics sheet for automatic emotion recognition and sentiment analysis," *Computational Linguistics*, vol. 48, no. 2, pp. 239–278, 2022.

[38] F. M. Plaza-del Arco and D. Nozza, "Respectful or toxic? using zero-shot learning with language models to detect hate speech," in *The 7th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics, 2023.

[39] F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "A multi-task learning approach to hate speech detection leveraging sentiment analysis," *IEEE Access*, vol. 9, pp. 112 478–112 489, 2021.

[40] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Detecting misogyny and xenophobia in Spanish tweets using language technologies," *ACM Transactions on Internet Technology*, vol. 20, no. 2, pp. 1–19, 2020.

[41] F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," *Expert Systems with Applications*, vol. 166, p. 114120, 2021.

[42] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang *et al.*, "AI alignment: A comprehensive survey," *arXiv preprint arXiv:2310.19852*, 2024.

[43] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[44] F.-J. Rodrigo-Ginés, J. Parra-Arnau, W. Meng, D. Rebollo-Monedero, and J. Forné, "PrivacySearch: An end-user and query generalization tool for privacy enhancement in web search," in *International Conference on Network and System Security*. Springer, 2018, pp. 312–327.

[45] F.-J. Rodrigo-Ginés, L. Plaza, and J. Carrillo-de Albornoz, "UNEDMediaBiasTeam@SemEval-2023 task 3: Can we detect persuasive techniques transferring knowledge from media bias detection?" in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics, 2023.

[46] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[47] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 1650–1659.

[48] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: an interdisciplinary literature review," *International Journal on Digital Libraries*, vol. 20, no. 4, pp. 391–415, 2019.

[49] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, pp. 1–15, 2020.