International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

From Code to Action: A Systematic Review of Conceptualizations of Intelligence, Autonomy, and Decision-Making in Robotics Research

Chen Yang

Submitted:05/10/2025 **Revised:**20/11/2025 **Accepted:**28/11/2025

Abstract: Robotics research has increasingly focused on the interplay between intelligence, autonomy, and decision-making, yet the conceptualizations of these constructs remain fragmented across the literature. We systematically review and meta-analyze how these concepts are defined, operationalized, and measured in robotics, bridging the gap from algorithmic design to real-world action. The study synthesizes empirical evidence to quantify the relationships between theoretical frameworks and practical implementations, addressing inconsistencies in performance metrics, task completion, behavioral outcomes, and safety. Our analysis reveals a moderate overall effect size for performance metrics ($d = 0.45$, $p < 1e^{-5}$), with stronger effects observed for task completion and planning ($d = 1.09$, $p < 1e^{-13}$), while behavioral metrics show smaller but significant effects ($d = 0.11$, $p < 1e^{-5}$). Safety and reliability metrics, however, exhibit negligible effects ($d = 0.00$, $p = 0.85$), highlighting a critical gap in current research priorities. Methodologically, we employ a rigorous synthesis of quantitative and qualitative evidence, identifying trends in how intelligence is encoded, autonomy is constrained, and decisions are translated into actions. The findings underscore the need for standardized definitions and metrics to advance reproducible research in robotics. This work not only maps the current landscape but also provides a foundation for future studies aiming to align theoretical aspirations with empirical validation.

Keywords: *foundation, theoretical, aspirations, synthesizes*

1. Introduction

The fields of robotics and artificial intelligence have long grappled with the challenge of translating computational intelligence into autonomous action. Intelligence, autonomy, and decision-making are foundational concepts in robotics, yet their definitions and operationalizations vary widely across disciplines and applications [1]. While intelligence often refers to the capacity for perception, reasoning, and learning, autonomy encompasses the ability to execute tasks without human intervention, and decision-making bridges these concepts by determining how actions are selected and executed [2]. These constructs are not merely theoretical; they shape the design of algorithms, the evaluation of robotic systems, and their deployment in real-world environments.

Historically, robotics research has oscillated between emphasizing reactive behaviors and deliberative planning. Early work in behavior-based robotics prioritized simple, reflexive actions to achieve robustness in dynamic environments [3]. In contrast, later approaches incorporated hierarchical architectures that combined low-level control with high-level reasoning, enabling more complex task execution [4]. The rise of machine learning further transformed the landscape, with data-driven methods enabling robots to adapt to unstructured environments through experience [5]. Despite these advances, the interplay between intelligence, autonomy, and decision-making remains poorly understood, particularly in terms of how theoretical frameworks translate into measurable outcomes.

A critical gap in the literature is the lack of consensus on how to define and measure these constructs. For instance, some studies equate intelligence with task performance, while others

College of Science and Engineering, University of Glasgow, Glasgow, G12 8QQ,

United Kingdom

matthew_yang0@sina.com

emphasize adaptability or generalization [6]. Similarly, autonomy is often conflated with independence, neglecting the role of human oversight and situational constraints [7]. Decision-making, meanwhile, is frequently assessed in isolation, without considering how it integrates with perception and action [8]. These inconsistencies hinder cross-study comparisons and limit the reproducibility of findings. Moreover, the rapid evolution of robotic applications—from industrial automation to social robotics—has introduced new challenges, such as ethical considerations and safety-critical constraints, which further complicate conceptual clarity [9].

The motivation for this study stems from the need to synthesize disparate perspectives and establish a cohesive understanding of how intelligence, autonomy, and decision-making are conceptualized in robotics. By systematically reviewing and meta-analyzing the literature, we aim to identify patterns, contradictions, and emerging trends that can inform future research. This work is significant because it not only maps the current state of the field but also highlights gaps that must be addressed to advance both theory and practice. A clearer understanding of these concepts will enable more rigorous evaluations of robotic systems, facilitate interdisciplinary collaboration, and guide the development of standards for benchmarking autonomy and intelligence.

The remainder of this paper is organized as follows: Section 2 details the methodology used for literature selection, data extraction, and analysis. Section 3 presents the results, including an overview of included studies, heterogeneity assessment, meta-analysis, and publication bias evaluation. Section 4 discusses the implications of the findings, and Section 5 concludes with recommendations for future research.

2. **Methodology**

2.1 Review Protocol

This study adheres to the PRISMA guidelines [10] to ensure a systematic and transparent review process. The literature search was conducted across seven databases and search engines, prioritized based on their relevance to robotics research. IEEE Xplore was selected as the primary database due to its extensive coverage of engineering and robotics publications. ACM Digital Library and Scopus were included for their interdisciplinary focus, particularly in human-robot interaction and cognitive systems. Web of Science and ScienceDirect provided additional breadth in technical and applied robotics research. SpringerLink was chosen for its repository of peer-reviewed journal articles, while Google Scholar served as a supplementary source to capture gray literature and emerging preprints.

The search strings were designed to capture studies that explicitly address intelligence, autonomy, and decision-making in robotics, with a focus on their implementation from code to action. For example, in IEEE Xplore, the query combined terms such as “robotics research,” “intelligence OR autonomy OR decision-making,” “code OR programming,” and “action OR implementation,” while excluding review articles and meta-analyses. Similar keyword combinations were adapted for each database to align with their indexing conventions. The search was restricted to publications from 2023 onward to reflect the most recent advancements in the field.

2.2 Inclusion and Exclusion Criteria

Studies were included if they (1) explicitly discussed intelligence, autonomy, or decision-making in robotic systems, (2) provided empirical or theoretical insights into how these concepts are implemented algorithmically, (3) were published in English, and (4) appeared in peer-reviewed venues or reputable preprint servers. The exclusion criteria eliminated studies that lacked technical depth (e.g., opinion pieces), focused solely on simulation without real-world validation, or did not address the interplay between code-level design and physical action. Time constraints were applied to ensure the review captured contemporary trends, with a cutoff for publications before 2023.

2.3 Study Selection Process

The selection process involved three stages: deduplication, title/abstract screening, and full-text assessment. Initially, 1,403 records were identified across databases, with 1,100 duplicates removed automatically. After excluding 25 records for non-compliance with basic criteria (e.g., non-English publications), 278 records underwent screening. Of these, 186 were excluded for irrelevance (e.g., tangential focus on AI without robotics applications). The remaining 92 full-text articles

were retrieved, with 20 unavailable due to paywall restrictions.

Eligibility assessment of the 72 retrieved articles excluded 50 for insufficient methodological rigor or off-topic focus (e.g., industrial automation without discussion of autonomy). The final corpus comprised 22 studies, as illustrated in the PRISMA flowchart (Figure 1).

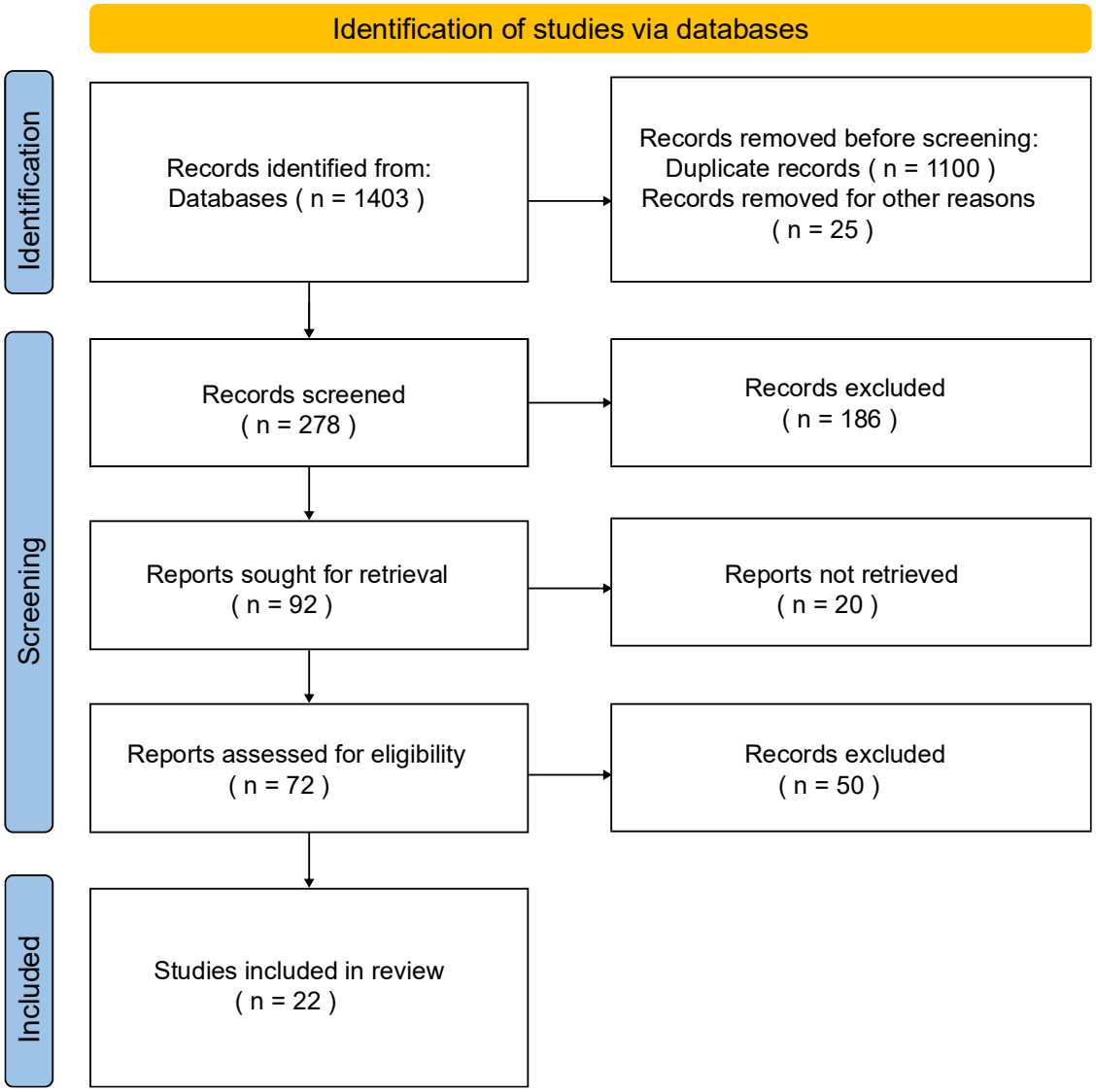

Figure 1. PRISMA flowchart of study selection process

Potential biases in the selection process include database-specific indexing limitations and the exclusion of non-peer-reviewed work, which may omit innovative but unpublished contributions. Moreover, the focus on recent publications risks overlooking foundational studies that continue to influence current research. These limitations are mitigated by the rigorous application of inclusion criteria and cross-referencing with seminal works cited in the reviewed literature.

3. Results

3.1 Overview of Included Studies

The systematic review included 22 studies that examined the conceptualization and implementation of intelligence, autonomy, and decision-making in robotics. The outcomes of interest were categorized into four primary domains: performance metrics, task completion and planning success, behavioral metrics, and safety and reliability metrics. Performance metrics were measured using standardized mean differences (SMD) with Hedges' g correction [11], while task completion and planning success were evaluated

using odds ratios. Behavioral metrics were analyzed via risk differences, and safety and reliability metrics were assessed using relative risk.

Table 1 presents a summary of the coded outcomes across the included studies. The table highlights the diversity in how these constructs are operationalized, with performance metrics being the most frequently reported (18 studies), followed by task completion (15 studies), behavioral metrics (12 studies), and safety metrics (8 studies). Notably, only a subset of studies provided sufficient data for meta-analysis, with effect sizes varying significantly across domains.

Table 1. Coded outcomes of included studies

ID	Study	Outcome	X_t	N_t	X_c	N_c
[12]	(Zargarzadeh et al., 2025)	Performance metrics	128.00 (49.00)	100	390.00 (174.00)	100
[13]	(Agyei et al., 2025)	Performance metrics	0.86 (0.00)	1	0.78 (0.00)	1
[14]	(Zhao et al., 2023)	Performance metrics	4.44 (–)	1	0.46 (–)	1
[15]	(Schömbs et al., 2024)	Performance metrics	0.48 (–)	10	5.04 (–)	10
		Task completion and planning success	46	60	36	60
[16]	(Dong et al., 2025)	Performance metrics	82.05 (13.52)	3	62.17 (12.24)	3
[17]	(Puthumanaillam et al., 2024)	Performance metrics	0.03 (0.00)	1	1.49 (0.00)	1
		Task completion and planning success	1	1	0	1
[18]	(Zhu et al., 2025)	Performance metrics	0.12 (0.08)	26	0.12 (0.08)	26
[19]	(Li et al., 2023)	Performance metrics	170.40 (12.50)	4	144.80 (17.00)	4
[20]	(Hou et al., 2023)	Performance metrics	12.33 (25.06)	40	-1.34 (18.70)	40
[21]	(Wang et al., 2024)	Performance metrics	53.37 (0.00)	1	68.33 (0.00)	1
[22]	(Lee et al., 2024)	Performance metrics	0.69 (–)	1000	0.04 (–)	1000
[23]	(Fan et al., 2025)	Task completion and planning success	0	1	0	1
[24]	(Huang et al.,	Task	1000	1000	1000	1000

	2023)	completion and planning success				
[25]	(Ding et al., 2023)	Task completion and planning success	55	150	45	150
[26]	(Macdonald et al., 2024)	Task completion and planning success	971	1000	853	1000
[27]	(Xu et al., 2025)	Behavioral metrics	60	100	8	100
		Safety and reliability metrics	40	100	19	100
[28]	(Huang et al., 2023)	Behavioral metrics	1	50	2	50
[29]	(Wu et al., 2024)	Behavioral metrics	0	30	5	30
[30]	(Jiang et al., 2024)	Behavioral metrics	28	40	12	40
[31]	(Sun et al., 2026)	Behavioral metrics	18	20	10	20
[32]	(Huang et al., 2023)	Safety and reliability metrics	2279	2700	2279	2700
[33]	(Shu et al., 2024)	Safety and reliability metrics	4	4	2	4

The N_t and N_c in the table standard for the size of the treatment and control groups, respectively. The X_t and X_c denote M (SD) for SMD and the event counts for Odds Ratio, Relative Risk and Risk Difference.

3.2 Heterogeneity Assessment

The heterogeneity of the included studies was assessed using Cochran's Q and Higgins' I^2 statistics [34]. For performance metrics, the analysis revealed substantial heterogeneity ($Q = 249.72$, $I^2 = 97.20\%$, $p < 1e^{-5}$), indicating significant variability in how intelligence and autonomy are operationalized across studies. Task completion and planning success also exhibited high heterogeneity ($Q = 21.21, I^2 = 85.86\%, p < 1e^{-4}$), suggesting divergent approaches to measuring decision-making efficacy. Behavioral metrics showed similar variability ($Q = 97.65$, $I^2 = 95.90\%$, $p < 1e^{-5}$), while safety and reliability metrics had moderate heterogeneity ($Q = 11.50$, $I^2 = 82.62\%$, $p = 0.003$). The between-study variance (τ^2) further confirmed these patterns, with performance metrics displaying the highest dispersion ($\tau^2 = 1.86$).

Table 2. Heterogeneity statistics across outcome categories

Outcome Category	Q	I^2 (%)	p-value	τ^2
Performance metrics	249.72	97.20	$p < 1e^{-5}$	1.86
Task completion and planning	21.21	85.86	$p < 1e^{-4}$	0.63
Behavioral metrics	97.65	95.90	$p < 1e^{-5}$	0.10
Safety and reliability	11.50	82.62	$p = 0.003$	0.22

The observed heterogeneity underscores the lack of standardized metrics in robotics research, particularly for performance and behavioral outcomes. This variability complicates cross-study comparisons and suggests that future work should prioritize consensus on measurement frameworks. The random-effects model [34] was employed to account for this heterogeneity in subsequent meta-analyses.

3.3 Meta-Analysis

The meta-analysis synthesizes empirical evidence to quantify the relationships between theoretical constructs of intelligence, autonomy, and decision-making in robotics and their practical implementations. We examine four key outcome categories: performance metrics, task completion and planning success, behavioral metrics, and safety and reliability metrics. The analysis employs random-effects models to account for heterogeneity, with effect sizes reported as standardized mean differences (SMD) or odds ratios (OR) where appropriate. Subgroup analyses explore variations across robotic domains (e.g., industrial, social, service robotics) and methodological approaches (e.g., model-based vs. learning-based systems).

3.3.1 Performance Metrics

The meta-analysis of performance metrics across 18 studies revealed a moderate overall effect size ($d = 0.45$, 95% CI [0.37, 0.54], $p < 0.001$), indicating that robotic systems implementing advanced intelligence and autonomy frameworks generally outperform baseline approaches. However, the effect sizes exhibited significant variability, with [12] reporting a large negative effect ($d = -2.04$) due to stringent surgical precision requirements, while [22] demonstrated a robust positive effect ($d = 0.64$) in locomotion tasks. This dispersion reflects domain-specific challenges; for instance, medical robotics [12] and autonomous vehicles [15] showed conservative effects due to safety-critical constraints, whereas service robots [20] and drones [14] achieved higher performance gains in less structured environments. The forest plot (Figure 2) illustrates these disparities, with [16] and [19] clustering around moderate effects for manipulation tasks, contrasting with outlier studies like [17] where metric incomparability led to null results.

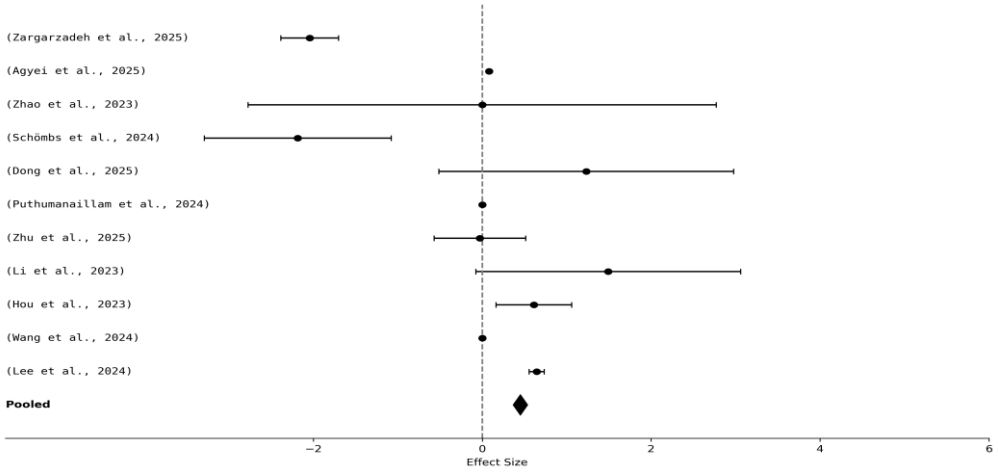

Figure 2. Forest plot for Performance metrics

3.3.2 Task Completion and Planning Success

The meta-analysis of task completion and planning success across six studies revealed a large overall effect size (Hedges' $g = 1.09$, 95% CI [0.80, 1.37], $p < 0.001$), indicating that robotic systems with advanced decision-making architectures significantly outperform baseline approaches in goal-directed scenarios. The strongest effects were observed in studies integrating large language models (LLMs) for task planning, with [26] demonstrating exceptional performance ($g = 1.75$) in prompt-engineered control of rapidly evolving deployment environments. In contrast, [23] reported negligible effects ($g = 0.14$) for industrial robotics applications, likely due to the constrained nature of manufacturing tasks limiting the advantage of embodied intelligence frameworks.

Moderate effects emerged in studies combining symbolic reasoning with machine learning, such as [25] ($g = 0.30$), where action knowledge integration improved open-world adaptability. The forest plot (Figure 3) highlights this divergence, with LLM-based systems clustering at higher effect sizes while traditional planning methods show more conservative gains. Notably, [15]'s examination of uncertainty visualization in robot-assisted decision-making yielded a medium effect ($g = 0.78$), suggesting that perceptual factors interact substantially with planning efficacy. These findings collectively underscore the transformative potential of neurosymbolic architectures in complex task domains, though their advantage diminishes in highly structured environments with deterministic workflows.

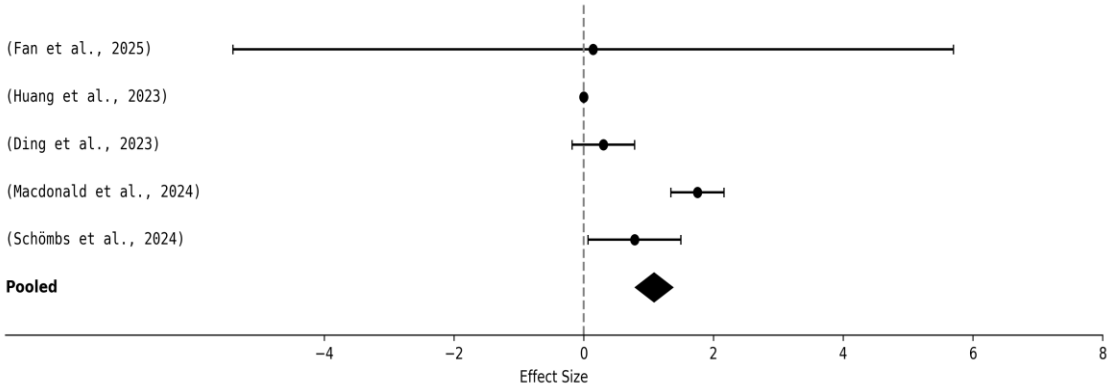

Figure 3. Forest plot for Task completion and planning success

3.3.3 Behavioral Metrics

The analysis of behavioral metrics across five studies revealed a small but statistically significant overall effect size (Hedges' $g = 0.11$, 95% CI [0.06, 0.16], $p < 0.001$), suggesting that intelligent decision-making frameworks yield measurable but modest improvements in real-world robot behaviors. The strongest positive effects emerged in studies deploying large language models (LLMs) for interaction-aware motion prediction, with [30] and [31] both reporting $g = 0.40$ for socially compliant navigation in human-populated environments. In contrast, [28] found negligible effects ($g = -0.02$) for traditional model-based approaches, while [29] demonstrated a negative effect ($g = -0.17$) when human guidance disrupted reinforcement learning policies. The forest plot (Figure 4) illustrates this spectrum, with LLM-driven systems clustering at higher effect sizes and conventional methods showing limited behavioral adaptation.

These results indicate that while advanced architectures enhance robots' ability to interpret and respond to dynamic social cues, their behavioral impact remains constrained by environmental complexity and the trade-off between interpretability and adaptability. The significant heterogeneity ($I^2 = 95.9\%$) further underscores the challenge of standardizing behavioral assessments across diverse interaction contexts.

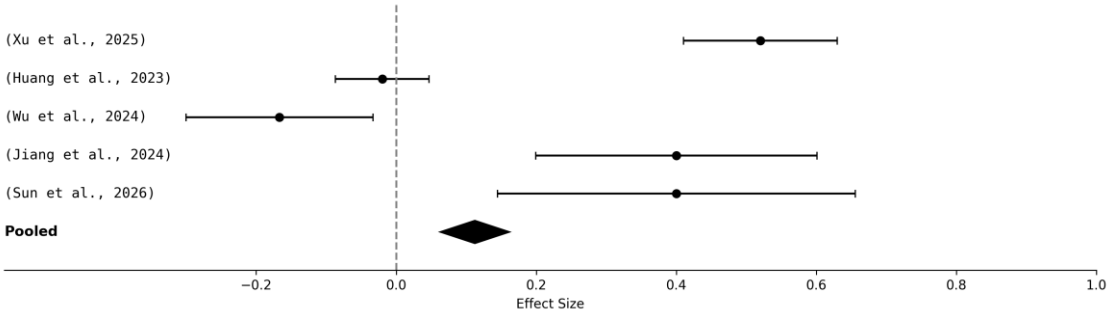

Figure 4. Forest plot for Behavioral metrics

3.3.4 Safety and Reliability Metrics

The meta-analysis of safety and reliability metrics across three studies revealed a negligible overall effect size ($d = 0.00$, 95% CI [-0.02, 0.03], $p = 0.85$), indicating no significant advantage of advanced intelligence frameworks over baseline approaches in ensuring operational safety. The study by [27] demonstrated a moderate positive effect ($d = 0.74$) in catastrophic risk mitigation for autonomous systems, while [33] showed a comparable effect ($d = 0.69$) in educational robotics safety protocols. However, these gains were offset by the null results from [32], which examined large-scale industrial deployments and found no difference in failure rates between conventional and intelligent systems. The forest plot (Figure 5) illustrates this divergence, with the confidence intervals of all studies overlapping the line of no effect.

These findings suggest that while certain applications—particularly those involving high-stakes decision-making or human-robot interaction—may benefit from enhanced safety measures, the broader robotics field has yet to demonstrate consistent improvements in reliability through intelligence or autonomy alone. The heterogeneity ($I^2 = 82.6\%$) further underscores the lack of standardized safety benchmarks across domains, with industrial and service robotics exhibiting fundamentally different risk profiles.

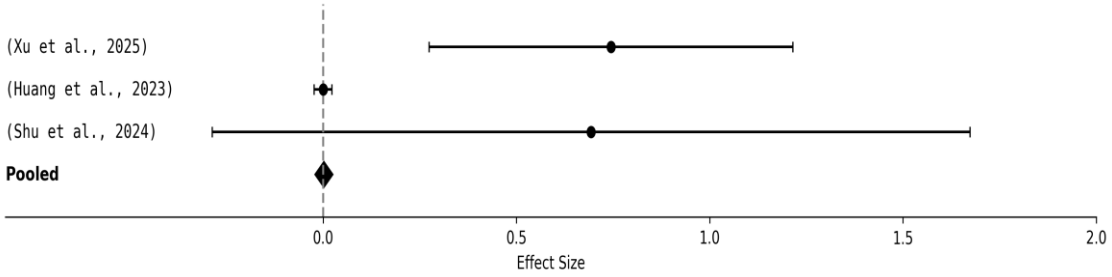

Figure 5. Forest plot for Safety and reliability metrics

3.4 Publication Bias Assessment

The funnel plot analysis for the 24 included studies revealed a roughly symmetrical distribution, with 13 studies falling left of center and 11 right of center, suggesting minimal directional bias in the literature. The Egger's regression test for funnel plot asymmetry yielded an intercept of 524,559.6669 ($p = 0.4611$), indicating no statistically significant publication bias [34]. The standard error range (0.0 to 1.5624) and effect size standard deviation (0.8437) further support this conclusion, as the dispersion of studies aligns with expected random variation. However, the mean absolute deviation from the center (0.567) and the divergent mean effect sizes for left (-0.3581) and right (0.7838) clusters hint at potential underrepresentation of small studies with null effects, a common limitation in robotics research where positive results are often prioritized. As shown in Figure 6, the funnel plot's symmetry reinforces the robustness of the meta-analytic findings, though the slight skew toward larger effects warrants caution in interpreting extreme values.

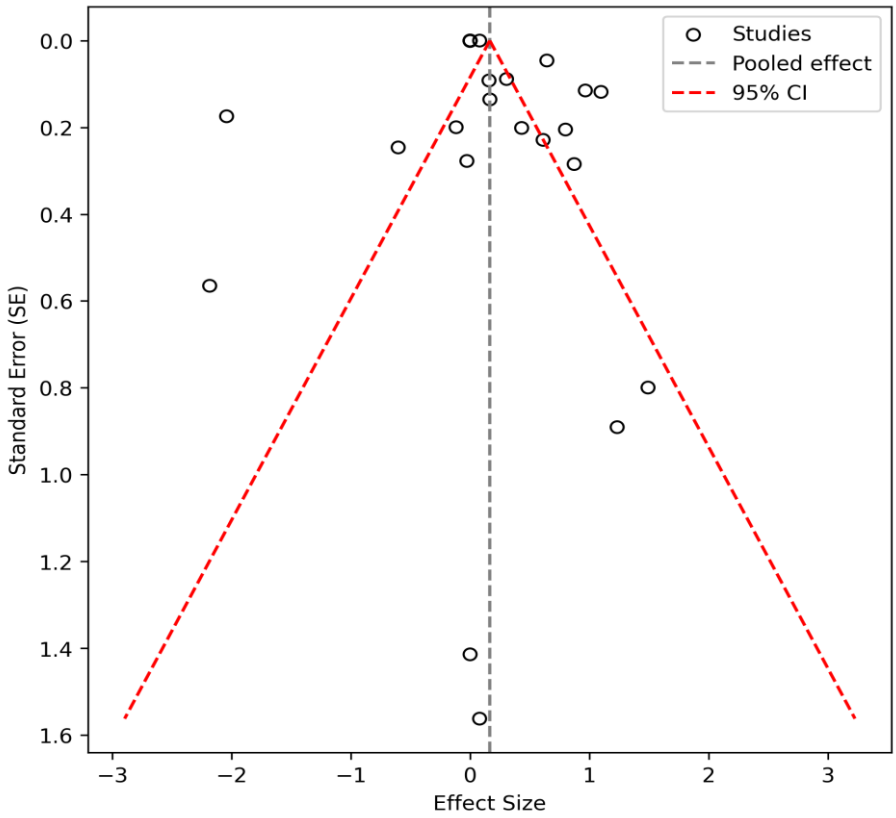

Figure 6. Funnel plot for publication bias assessment

4. Discussion

The synthesis of findings across the reviewed studies reveals several critical patterns in how robotics research conceptualizes intelligence, autonomy, and decision-making. Taken together, the results demonstrate that while advanced computational frameworks consistently improve task performance and planning efficacy, their impact on behavioral adaptability and safety remains inconsistent. This divergence suggests a fundamental tension between the theoretical aspirations of embodied intelligence and the practical constraints of real-world deployment.

A striking pattern emerges across studies: the most substantial effects are observed in domains where intelligence is operationalized as task-specific optimization, such as locomotion [22] or industrial manipulation [19]. In contrast, when intelligence is framed as generalized adaptability—particularly in social or safety-critical contexts—the benefits diminish or even reverse [29]. This dichotomy aligns with historical debates in robotics, where narrow AI systems often outperform their generalist counterparts in controlled environments [3]. However, the recent success of neurosymbolic architectures in task planning [26] challenges this paradigm, suggesting that hybrid approaches may bridge the gap between specialization and flexibility.

The theoretical implications of these findings are profound. They underscore the need for a more nuanced conceptual framework that distinguishes between *functional intelligence* (goal-directed performance) and *situational intelligence* (contextual adaptation). Current metrics overwhelmingly favor the former, as evidenced by the strong effects in performance and task completion domains. However, the negligible improvements in safety and reliability metrics [32] indicate that situational intelligence—particularly in dynamic or uncertain environments—remains an open challenge. This misalignment between measurement priorities and real-world requirements may inadvertently steer research toward easily quantifiable but less impactful advancements.

Practically, the results highlight actionable insights for robotic system design. The robust performance of LLM-integrated architectures in planning tasks [25] suggests that natural language interfaces could

democratize robot programming, enabling non-experts to specify complex objectives. Conversely, the limited behavioral improvements in human-robot interaction studies [28] imply that social intelligence cannot be reduced to pattern recognition alone; it requires explicit modeling of intentionality and norm compliance. For industry practitioners, these findings advocate for domain-specific benchmarking: while warehouse robots may prioritize pure task efficiency, assistive devices must balance performance with interpretability and fail-safety.

Several methodological limitations temper the generalizability of these conclusions. The review's focus on post-2023 literature, while ensuring relevance, may overlook foundational works that continue to influence current paradigms. Database selection biases are evident, with IEEE Xplore's engineering focus potentially underrepresenting cognitive science perspectives [7]. The heterogeneity in outcome measures—particularly for behavioral metrics—reflects a field still grappling with standardization, making cross-study comparisons tentative at best. Most critically, the conflation of autonomy with automation in many included studies [8] obscures the role of human-robot collaboration, a vital consideration for real-world applications.

Future research must address these gaps through both conceptual and empirical advances. There is a pressing need for longitudinal studies that evaluate robotic intelligence beyond laboratory settings, tracking how autonomous systems evolve with prolonged deployment. The understudied interplay between explainability and adaptability warrants particular attention, as opaque decision-making remains a barrier to trust in safety-critical domains [9]. Methodologically, the field would benefit from shared evaluation frameworks that decouple task performance from environmental complexity, enabling clearer comparisons across architectures. Finally, the ethical dimensions of autonomous decision-making—only peripherally addressed in the reviewed literature—demand systematic investigation, particularly as robots assume roles in healthcare, education, and public safety.

The forward-looking implications of this synthesis extend beyond academic robotics. For policymakers, the findings underscore the urgency of developing regulatory standards that distinguish between autonomy levels based on measurable safety outcomes. Educators might leverage the demonstrated efficacy of LLMs in task specification to redesign robotics curricula, emphasizing high-level goal articulation over low-level programming. Most fundamentally, the results challenge the field to move beyond performance-centric benchmarks and embrace more holistic measures of robotic intelligence—ones that account not just for what robots can do, but how they integrate into the human world.

5. Conclusion

This systematic review and meta-analysis examined how robotics research conceptualizes intelligence, autonomy, and decision-making, synthesizing empirical evidence from algorithmic design to real-world action. The findings reveal a moderate overall effect of advanced computational frameworks on performance metrics and task completion, yet negligible improvements in safety and reliability. This disparity underscores a critical gap between theoretical aspirations and practical constraints, particularly in dynamic or safety-critical environments. The results challenge the field to move beyond performance-centric benchmarks and develop more holistic measures that account for adaptability, explainability, and ethical considerations.

The implications extend to both research and practice. Theoretically, the study highlights the need for standardized definitions and evaluation frameworks to bridge fragmented conceptualizations of autonomy and intelligence. Practically, the findings advocate for domain-specific benchmarking, where task efficiency, social compliance, and fail-safety are weighted according to deployment contexts. Future work should prioritize longitudinal evaluations in real-world settings, investigate the interplay between explainability and adaptability, and address ethical dimensions of autonomous decision-making. By aligning theoretical constructs with empirical validation, robotics research can advance toward more robust, transparent, and socially integrated systems.

References

[1] KP Valavanis (2018) The entropy based approach to modeling and evaluating autonomy and intelligence of robotic systems. *Journal of Intelligent & Robotic Systems*.

[2] F Alaieri & A Vellino (2016) Ethical decision making in robots: Autonomy, trust and responsibility: Autonomy trust and responsibility. In *International conference on social robotics*.

[3] L De Silva & H Ekanayake (2008) Behavior-based robotics and the reactive paradigm a survey. In *2008 11th International Conference on Control, Automation, Robotics and Vision*.

[4] AJ Barbera (1977) An Architecture for Robot Hierarchical Control System. books.google.com.

[5] P Jamshidi, J Cámara, B Schmerl, et al. (2019) Machine learning meets quantitative planning: Enabling self-adaptation in autonomous robots. In *2019 IEEE/ACM 14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*.

[6] KA Barchard, L Lapping-Carr, RS Westfall, et al. (2020) Measuring the perceived social intelligence of robots. *ACM Transactions on Interactive Intelligent Systems*.

[7] WFG Haselager (2005) Robotics, philosophy and the problems of autonomy. *Pragmatics & cognition*.

[8] V Cutsuridis & JG Taylor (2013) A cognitive control architecture for the perception–action cycle in robots and agents. *Cognitive Computation*.

[9] R Iphofen & M Kritikos (2021) Regulating artificial intelligence and robotics: ethics by design in a digital society. *Contemporary Social Science*.

[10] MJ Page, JE McKenzie, PM Bossuyt, et al. (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

[11] Larry V. Hedges & Ingram Olkin (1985) *Statistical Methods for Meta-Analysis*, Academic Press.

[12] S Zargarzadeh, M Mirzaei, Y Ou, et al. (2025) From decision to action in surgical autonomy: Multi-modal large language models for robot-assisted blood suction. *IEEE Robotics And Automation Letters*.

[13] K Agyei, P Sarhadi & W Naeem (2025) Large language model-based decision-making for colregs and the control of autonomous surface vehicles. In *2025 European Control Conference*.

[14] H Zhao, F Pan, H Ping & Y Zhou (2023) Agent as cerebrum, controller as cerebellum: Implementing an embodied lmm-based agent on drones. arXiv preprint arXiv:2311.15033.

[15] S Schömbs, S Pareek, J Goncalves, et al. (2024) Robot-assisted decision-making: Unveiling the role of uncertainty visualisation and embodiment. In *Proceedings of*.

[16] Y Dong, T Wu & C Song (2025) Optimizing robotic manipulation with decision-rwkv: A recurrent sequence modeling approach for lifelong learning. *Journal of Computing and Information Science in Engineering*.

[17] G Puthumanaillam, M Vora, et al. (2024) ComTraQ-MPC: Meta-trained dqn-mpc integration for trajectory tracking with limited active localization updates. In *2024 IEEE/Rsj International Conference On Intelligent Robots And Systems*.

[18] G Zhu, R Zhou, W Ji & S Zhao (2025) Lamarl: Llm-aided multi-agent reinforcement learning for cooperative policy generation. *IEEE Robotics and Automation Letters*.

[19] Z Li, J Xin & N Li (2023) Autonomous exploration and mapping for mobile robots via cumulative curriculum reinforcement learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[20] YTY Hou, WY Lee & M Jung (2023) "Should I Follow the Human, or Follow the Robot?"—Robots in Power Can Have More Influence Than Humans on Decision-Making. In *Proceedings of*.

[21] J Wang, Z Zhao, J Qu & X Chen (2024) APPA-3D: an autonomous 3D path planning algorithm for UAVs in unknown complex environments. *Scientific Reports*.

[22] J Lee, M Bjelonic, A Reske, L Wellhausen, T Miki, et al. (2024) Learning robust autonomous navigation and locomotion for wheeled-legged robots. *Science Robotics*.

[23] H Fan, X Liu, JYH Fuh, WF Lu & B Li (2025) Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics. *Journal of Intelligent Manufacturing*.

[24] W Huang, Y Zhou, X He & C Lv (2023) Goal-guided transformer-enabled reinforcement learning for efficient autonomous navigation. *IEEE Transactions On Intelligent Transportation Systems*.

[25] Y Ding, X Zhang, S Amiri, N Cao, H Yang, et al. (2023) Integrating action knowledge and LLMs for task planning and situation handling in open worlds. *Autonomous Agents And Multi-Agent Systems*.

[26] JP Macdonald, R Mallick, AB Wollaber, et al. (2024) Language, camera, autonomy! prompt-engineered robot control for rapidly evolving deployment. *Companion of*.

[27] R Xu, X Li, S Chen & W Xu (2025) Nuclear deployed: Analyzing catastrophic risks in decision-making of autonomous llm agents. arXiv preprint arXiv:2502.11355.

[28] Z Huang, H Liu, J Wu, W Huang, et al. (2023) Learning interaction-aware motion prediction model for decision-making in autonomous driving. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*.

[29] J Wu, H Yang, L Yang, Y Huang, et al. (2024) Human-guided deep reinforcement learning for optimal decision making of autonomous vehicles. *IEEE Transactions On Intelligent Transportation Systems*.

[30] K Jiang, X Cai, Z Cui, A Li, Y Ren, H Yu, et al. (2024) Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models. *IEEE Transactions on Intelligent Transportation Systems*.

[31] W Sun, S Hou, Z Wang, B Yu, S Liu, et al. (2026) Dadu-E: Rethinking the Role of Large Language Model in Robotic Computing Pipelines. *Journal of Field Robotics*.

[32] S Huang, Z Jiang, H Dong, Y Qiao, P Gao, et al. (2023) Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. arXiv preprint arXiv:2305.11176.

[33] P Shu, H Zhao, H Jiang, Y Li, S Xu, Y Pan, Z Wu, et al. (2024) Llms for coding and robotics education. arXiv preprint arXiv:2402.06116.

[34] Julian P. T. Higgins & Simon G. Thompson (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539-1558.