Survey on Security and Safety in Humanoid Robotics

Mingxiao Wei[1], Yubin Ma[2], Mingyang Qin[2], Haohong Wang[3], and Jian Ren[4]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Department of Statistics, University of California, Davis, CA 95616, USA
[3]TCL Research American, San Jose, CA 95110 USA
[4]Department of ECE, Michigan State University, East Lansing, MI 48824, USA.

***Abstract*—As humanoid robots become increasingly integrated into healthcare, domestic assistance, and public spaces, ensuring their security and safety has become paramount. This survey synthesizes current research on ethical frameworks, security vulnerabilities, and safety considerations specific to humanoid robotics systems. We examine the evolution from early conceptual frameworks to contemporary approaches addressing adversarial threats, privacy concerns, and human-robot interaction safety. The survey identifies critical gaps in current frameworks and proposes directions for developing comprehensive, actionable security and safety guidelines for humanoid robots operating in human-centered environments.**

***Index Terms*—Humanoid robotics, security, safety, privacy, adversarial attacks, human-robot interaction**

I. Introduction

Humanoid robots represent a unique convergence of artificial intelligence, physical embodiment, and social interaction capabilities. Unlike industrial robots or autonomous vehicles, humanoid robots are designed to operate in intimate proximity to humans, often serving vulnerable populations including children, elderly individuals, and patients in healthcare settings. This proximity introduces distinctive security and safety challenges that extend beyond traditional robotics concerns.

The rapid advancement of humanoid robotics over the past two decades has outpaced the development of comprehensive security and safety frameworks. Early ethical guidelines, exemplified by Isaac Asimov's fictional "Three Laws of Robotics," provided cultural touchstones but lacked practical applicability to real-world socio-technical systems [1]. Contemporary research has shifted toward responsible robotics frameworks that emphasize distributed accountability, stakeholder participation, and context-sensitive design [2], [3].

This survey examines three interconnected dimensions of humanoid robotics security and safety. First, we explore ethical governance frameworks that define principles and accountability structures. Second, we analyze technical security considerations addressing vulnerabilities across system layers. Third, we investigate safety mechanisms ensuring physical and psychological harm prevention. Through this comprehensive examination, we identify critical gaps in existing approaches and propose directions for future research and development.

II. Evolution of Ethical Frameworks

A. From Asimov to Responsible Robotics

Murphy and Woods challenged Asimov's robot-centered ethical model, arguing that morality cannot be programmed solely within machines [1]. Real-world robots operate within socio-technical systems where responsibility spans designers, users, and institutions. Asimov's laws represent functional morality, neglecting human decision-making and accountability—already evident in autonomous military systems that override the "no harm" principle.

Their alternative framework places ethics on human agents. The first revised law prohibits deploying robots unless their design and operation meet the highest safety and ethical standards, supported by behavioral traceability tools [1]. The second requires context-aware interaction with humans, while the third stresses situated autonomy—robots maintaining independence yet allowing smooth control transfer. This shift redefines robotics ethics around system-level accountability and resilience rather than individual robot behavior.

B. Global Convergence on Ethical Principles

A review of 84 AI ethics guidelines found consistent emphasis on five principles—transparency, justice, non-maleficence, responsibility, and privacy [4]. Transparency dominates, while non-maleficence outweighs beneficence, reflecting a stronger focus on preventing harm. Yet interpretations and enforcement mechanisms vary widely. Most guidelines originate from developed nations and are produced by private and public organizations in nearly equal measure, raising concerns about global representation [4]. The results suggest that closer collaboration and standardized governance are needed to harmonize ethical practices.

C. Accountability Distribution

Tóth et al. outlined two dimensions of AI accountability: locus of morality (where ethical agency resides) and moral intensity (the contextual severity of consequences) [5]. Combining these yields four clusters—professional norms, business responsibility, inter-institutional normativity, and supra-territorial regulation—each reflecting increasing dispersion of accountability. Humanoid robots may traverse these clusters depending on function, from low-risk companions to high-stakes medical or defense systems. The framework aids policymakers in aligning responsibility with AI autonomy and evolving application contexts [5].

III. Security Considerations for Humanoid Robots

A. Extended Robotics Networks

Humanoid robots operate as components within what can be termed Extended Robotics Networks (ERNs). These com-

prehensive systems encompass the physical robot, detached sensors, storage systems, processing centers, communication networks, and teleoperation agents. Each layer introduces distinct vulnerabilities requiring targeted security measures. The concept of ERNs proves essential for understanding the complete attack surface of humanoid robotics systems, as security cannot be adequately addressed by focusing solely on the embodied robot itself.

ERNs contain many layers, including the sensors, physical body, communication networks, processing centers, artificial intelligence model, and database systems. They all have distinct vulnerabilities. A proper security guide for ethical and responsible robotics design must address each of the layers individually and offer instructions. Since these robots will be deployed for health and well-being assistance, disrupting their activities may be life or death in some scenarios.

B. Attack Surfaces and Threat Vectors

Humanoid robots face multi-layered threats. Physical attacks include jamming, spoofing, and sensor manipulation, while software vulnerabilities in libraries like OpenCV or TensorFlow pose additional risks [6], [7]. Hardware Trojans and side-channel exploits further endanger system integrity [8]. Defenses involve filtering, anomaly detection, sensor fusion, tamper-resistant design, secure hardware, and software verification.

Network attacks such as interception, man-in-the-middle, denial-of-service, and data leakage compromise privacy and functionality [6], [9]. Mitigation strategies include encryption, firewalls, TLS/DTLS, local processing, and reduced data exchange.

Software and AI attacks target training (data poisoning), inference (model inversion), and output manipulation, with social engineering via natural language interfaces posing additional risks [6], [8]. Defenses include adversarial training, input sanitization, privacy-preserving methods, formal verification, and regular patching. Firmware threats such as ransomware and trojans require layered software protection.

Authentication and authorization attacks exploit identity and access control weaknesses. Multi-factor and biometric verification, role-based access control, and cryptographic protocols help ensure only authorized entities operate robots [1], [8]. Evaluation frameworks assess the robustness of these mechanisms.

C. Privacy and Surveillance Concerns

Humanoid robots create unique privacy risks across sensor, device, network, cloud, and software layers due to cameras, microphones, and local storage [6], [9], [10]. Mitigations include sensor-by-demand architectures, tamper-resistant hardware, physical isolation, encryption, and air-gapped or local-processing systems for sensitive populations [6], [7], [9], [11].

Privacy is increasingly defined as user agency—the ability to dynamically control or revoke data usage [12]. Studies with older adults show preferences for robots that can delete data, ensure transparency, and provide empathetic interaction while safeguarding personal information [13]. Policies should specify retention, usage, and user-controlled deletion.

Privacy-enhancing technologies include data alteration (anonymization, differential privacy), shielding (cryptography, secure communication, blockchain), and system-level strategies (privacy-by-design, local storage, sensor minimization, access control) [6], [8], [12]. Additional methods address personalization risks via fingerprinting, obfuscation, self-destructing data, and statistical disclosure controls, collectively empowering users to manage privacy across the data lifecycle [6].

D. Security Solutions and Countermeasures

Cryptographic Solutions Cryptographic protocols authenticate users or devices, with symmetric methods favored for lightweight, energy-efficient humanoid robots. ROS communications can be secured using TLS and DTLS, providing fine-grained control over data publishing and subscription, while cryptography helps mitigate denial-of-service attacks.

Cloud cybersecurity enhancements include encrypted clustering (e.g., AES), cloud-edge hybrid systems for secure feedback control, Kerberos- and elliptic-curve-based authentication, and blockchain architectures for tamper-resistant, accountable connections.

While strong encryption addresses most communication vulnerabilities, physical attacks such as jamming remain a critical challenge, highlighting the need to balance security and operational resilience in humanoid robotics.

Intrusion Detection Systems Intrusion detection systems (IDS) enhance protection against security threats in humanoid robotics using architectural, statistical, and machine learning methods, achieving accuracies of 88–100% with low false positives [8]. Robotic Intrusion Prevention Systems (RIPS) integrate security and safety for cognitive robots, using event-driven actions, and boolean expressions to adapt system behavior based on threat detection [14]. De-escalation from serious alerts requires human oversight, ensuring critical incidents are managed safely.

RIPS handles message, graph, and external events with alert, set, exec, and trigger actions, configurable via model files that define system structure and parameters. Experiments on TIAGo robots confirmed that RIPS responds correctly to security changes while maintaining stable operation [14]. System modes enable integration of security with the overall cognitive architecture, making security a core system feature rather than an add-on.

Machine learning approaches—including k-nearest neighbors, naive Bayes, k-means, adaptive boosting, ensembles, and unsupervised techniques—support detection of evolving and zero-day attacks [8], [15]. ROS immunity solutions further harden systems with low overhead through robustness assessment, automatic rule generation, and distributed defenses, combining IDS with proactive protection strategies.

IV. Safety Considerations

A. Non-Maleficence as Core Principle

The Principle of Non-maleficence holds that moral agents must avoid causing direct harm to humans or other beings within their responsibility [16]. It prioritizes harm prevention over utility maximization and distinguishes active harm from harm by omission—an essential distinction for robots facing

ethically complex decisions. Originating in biomedical ethics, it provides a foundation for healthcare robotics, where autonomous systems increasingly affect patient wellbeing.

The Partnership Principle extends non-maleficence to human–robot collaboration [16], requiring that humans work only with robots that uphold their moral and legal standards. This principle ensures moral alignment, prevents accountability offloading, and supports coordination in professional contexts—especially in healthcare, where both humans and robots share responsibility for patient safety.

B. Implementation Challenges

Implementing non-maleficence in humanoid robots faces five categories of challenges [16].

Conceptual challenges center on defining the sphere of responsibility. The principle necessitates specifying what scope of harms a robot should consider and prevent, including both actions and inactions. This varies by context—a companion robot in a home has different responsibilities than a healthcare robot in a hospital, and the scope changes based on specific situations and vulnerability of nearby individuals.

Technical challenges involve developing a "harm ontology" [16] that enables robots to identify relevant features, predict harm sources, and assess both physical harms (collisions, injuries) and non-physical harms (privacy violations, emotional harm from psychological attachment). The comprehensiveness and accuracy of this ontology directly determine a robot's ability to avoid harm.

Moral challenges require robots to distinguish morally salient harms within context [16]. Robots may need to exert permissible localized harms (e.g., temporary discomfort during medical procedures) and determine responsibility for preventable harms they didn't directly cause. This requires sophisticated moral reasoning beyond rule-following to understand context, intention, and proportionality.

Political challenges arise because ethical tradeoffs should involve collective decision-making [16]. Decisions about acceptable risks and competing values require diverse stakeholder input rather than solely engineers or corporate decision-makers, necessitating appropriate governance structures and inclusive deliberation processes.

Explainability challenges relate to transparent harm detection mechanisms [16]. Determining whether robot-caused harms are systematic or random requires explainable harm ontologies incorporated from earliest design stages. Without explainability, diagnosing failures and assigning accountability becomes difficult.

C. Variable Autonomy and Safe Control Transfer

Variable autonomy—dynamic shifts between autonomy levels during operation—serves as a key safety strategy for humanoid robots [17]. Research reveals several implementation dimensions affecting safety outcomes.

Initiative determines who initiates autonomy changes: human initiative (operator-only), system initiative (automatic), or mixed initiative (collaborative) [17]. Specificity defines what changes: traded control shifts between fully manual and automatic extremes, while granular control adjusts functions along a continuous scale. Flexibility determines when changes occur: goal-oriented approaches pre-define changes, stimulus-driven approaches decide at runtime, and hybrid approaches combine both [17]. Triggers activate changes based on task aspects, operator states, system events, or environmental circumstances.

Research predominantly uses experimental, simulation, and field test methodologies, though mostly in contrived environments that fail to capture real-world complexity [17]. Evaluation measures include capability constructs (task performance) and collaboration constructs (interaction quality), but rely heavily on ad hoc measures rather than established tools like the NASA Task Load Index.

Researchers pursue variable autonomy to improve effectiveness, efficiency, and safety of human-robot teams by leveraging autonomous behaviors to reduce human stress while utilizing human strengths in complex environments [17]—particularly important for humanoid robots in unpredictable settings where rigid autonomy compromises safety or functionality.

D. Human-Robot Interaction Safety

Human-robot interaction (HRI) safety extends beyond physical harm prevention to cognitive, social, and informational domains. Physical safety reduces accident risks during interaction, while cognitive safety ensures understandable robot behavior and appropriately calibrated trust. Paradoxically, perfectly reliable robots may seem unrealistic, while error-prone robots can appear more believable, requiring balance based on task criticality and user expectations [6].

Social safety focuses on acceptance, avoiding dependence, and supporting human-human connections. Studies with older adults show companion robots can facilitate emotional discussion but are insufficient to alleviate loneliness, highlighting the need for robots to enhance, not replace, social relationships [13].

Informational safety protects against misinformation and manipulation, particularly given that large language models can produce plausible but false outputs. Hallucination prevention and fact-checking are crucial, especially for robots providing health-related advice [13].

Internet-connected robots face cybersecurity threats, requiring transparency in data handling. Human trust depends on robot reliability and design, task complexity, and individual user differences [6].

V. Application-Specific Considerations

A. Conversational Companion Robots for Older Adults

Thematic analysis revealed insights into shared expectations around interaction, personalization, privacy, information, and emotional support [13].

Active and Passive Listening: Active listening enables the robot to engage in conversation, understand context, and ask follow-up questions when the user is alone. LLMs can extract relevant dialogue and generate meaningful questions. Passive listening allows the robot to observe social environments quietly and later help the user reflect on group interactions.

Personalization: Participants emphasized personalization as understanding preferences and relationships, evolving over time like human-human interactions. Robots should ask ques-

tions early, provide context-based advice, and leverage lifelong learning via LLMs for personalized recommendations.

Privacy and Data Control: Users expect the robot to identify them, delete undesired data, and prevent leakage. Cloud-based computation may require data anonymization, while open-world learning could address challenges like user recognition bias and inaccuracies.

Information Retrieval: Robots should remind users of schedules, provide context-aware explanations, and fact-check to reduce reliance on phones. Credible, up-to-date information is essential, especially for health-related advice, requiring hallucination prevention and knowledge-base verification.

Emotional and Social Expectations: Participants value empathetic responses, cognitive assistance, and congruent emotional reflection. Emotional awareness via LLMs, multimodal affect recognition, appropriate vocal intonation, and mimicking expressions enhance likeability. Participants criticized current robots' limited facial expression and voice intonation richness.

B. Healthcare Robots

Healthcare applications introduce heightened safety stakes requiring specialized security and safety considerations [16]. Medical robots can inflict severe physical harm through incorrect actions such as inappropriate medication dosage. Patients, especially those with cognitive impairments, represent highly vulnerable populations requiring enhanced protections. Medical information carries special legal and ethical protections under regulations such as HIPAA in the United States and GDPR in Europe [8].

Security measures must address potential triggered maleficence—adversarial manipulation causing deliberate harm to patients. The 2015 Jeep security vulnerability, which resulted in $105 million in fines for Chrysler, illustrates the severe consequences of inadequate security in cyber-physical systems [6]. For healthcare robots, such vulnerabilities could be exploited to cause direct patient harm, making robust security essential rather than optional.

Regulatory requirements for medical devices include stringent certification and compliance processes [16]. Medical robotics must meet higher standards than consumer devices due to the potential for serious harm. This necessitates formal verification methods, extensive testing protocols, and ongoing monitoring systems. The challenge lies in balancing innovation with safety, ensuring that new capabilities do not introduce unacceptable risks.

C. Robots for Vulnerable Populations

Robots assisting dementia patients and other vulnerable populations require specialized security architectures [11]. Long Short-Term Memory models enable contextual understanding of interactions while maintaining continuity across sessions. IoT integration provides secure connection with monitoring systems and healthcare infrastructure, but introduces additional attack surfaces requiring careful management.

Physical security through tamper-resistant design prevents unauthorized access or manipulation [11]. Isolation strategies protect sensitive components through physical separation from potentially compromised elements. Behavioral monitoring detects anomalous patterns that may indicate security breaches or system malfunctions, but must be implemented with careful attention to dignity and privacy concerns.

The interaction between security requirements and the needs of vulnerable populations presents unique challenges. Systems must be secure enough to prevent exploitation while remaining accessible and non-threatening to users who may have limited technical understanding or cognitive impairments. Interface design must balance security controls with usability, ensuring that authentication mechanisms do not create barriers to legitimate use.

VI. Critical Gaps in Current Frameworks

A. Implementation Specificity

Although ethical principles show global convergence, current frameworks lack operational guidance [4], [6]. Security is often mentioned only briefly despite awareness of hacking, network, and physical risks. The gap between aspirational ethics and concrete implementation remains one of the main barriers to effective humanoid robotics governance.

B. Underrepresentation of Physical Security

Compared with privacy and transparency, physical security receives limited attention [6]. Few frameworks mandate tamper-resistance or address hardware-specific attack vectors such as sensor manipulation or Trojan insertion. Given humanoid robots' physical embodiment, standards ensuring hardware integrity are essential yet underdeveloped.

C. Limited Empirical Validation

Most frameworks are tested only in controlled lab environments using convenience samples rather than real-world users [17]. Such validation cannot capture the complexity, unpredictability, or adversarial pressures of actual deployment. Long-term, in-situ studies are needed to assess evolving risks and cumulative effects over time.

D. Measurement Inconsistency

Variable autonomy research often relies on ad hoc metrics instead of standardized tools like the NASA Task Load Index [17]. This inconsistency hampers cross-study comparability and knowledge accumulation. Establishing shared benchmarks for security, safety, ethics, and user experience would enable systematic evaluation.

E. Responsibility Diffusion

In distributed Extended Robotics Networks, accountability is fragmented among developers, manufacturers, service providers, and operators [5]. Machine learning's adaptive behavior and delayed harms further obscure liability. Clear accountability and traceability mechanisms are needed to prevent ethical and legal gaps.

F. Context Sensitivity Gap

Most frameworks apply universal principles despite varied ethical and technical demands across domestic, clinical, and public settings [13], [16], [17]. One-size-fits-all models risk being overly rigid or insufficiently protective. Context-sensitive approaches must align risk level, vulnerability, and cultural factors while maintaining overarching ethical coherence.

VII. Recommendations for Future Frameworks

Layered Security Architecture Comprehensive protection must span all ERN layers. Sensors use demand-based activation, fusion, and anti-spoofing defenses; devices rely on encryption, secure boot, and isolated execution; networks employ TLS, firewalls, and intrusion detection; clouds apply encryption, anonymization, and blockchain for auditability; AI models integrate adversarial training, differential privacy, and explainability; and applications enforce validation, secure APIs, and access control. Cross-layer mechanisms like unified RIPS ensure coordinated and traceable responses.

Privacy as Dynamic User Agency Privacy should empower users with real-time, contextual control over data use and retention. Transparency tools—dashboards, visualizations, and conversational interfaces—enable monitoring, while updated legal frameworks must guarantee enforceable data rights beyond GDPR.

Participatory, Stakeholder-Inclusive Design Security and safety frameworks must involve diverse users, caregivers, and ethicists from early stages. Continuous engagement, fair compensation, and accessible communication ensure inclusivity and genuine influence over design decisions.

Enforceable Standards and Graduated Compliance Robust regulations require mandatory baselines, domain-specific rules for high-risk contexts, certification, audits, and transparent incident reporting. Clear liability allocation and proportional compliance balance safety with innovation.

Integrated Safety–Security Design Unified design links security states with safety responses. Graceful degradation, alert escalation, and human takeover mechanisms ensure resilient operation under threat, supported by joint monitoring and response protocols.

Empirical Validation Frameworks demand long-term, real-world testing, adversarial evaluations, standardized metrics, and representative users to capture evolving trust and cumulative risks beyond lab settings.

Explainable Security and Safety Transparency through interpretable models, detailed logging, and tailored explanations fosters trust and accountability across all stakeholders—users, operators, regulators, and researchers.

VIII. Theoretical Foundations: Learning from Untrusted Data

Robust humanoid learning can tolerate corrupted inputs through list-decodable and semi-verified learning, which combine trustworthy anchors with cautious integration of untrusted data [18]. Adversarial and genuine samples are often separable via clustering, enabling anomaly detection. Using spectral bounds and convex objectives, robust optimization achieves accurate estimates even with limited clean data. These theories underpin spectral, clustering, and robust methods in humanoid robotics, balancing computational efficiency with resilience to manipulation.

IX. Conclusion

Humanoid robot security and safety demand integrated, enforceable frameworks combining ethics, technical safeguards, and stakeholder input. Guided by transparency, accountability, fairness, non-maleficence, and privacy, defenses span cryptography, intrusion detection, and adaptive safety controls. Remaining gaps include physical security, dynamic privacy, and real-world validation. Future systems should unify layered security, user-centered privacy, participatory design, and enforceable standards to ensure trustworthy, human-centered deployment.

References

[1] R. Murphy and D. D. Woods, "Beyond asimov: The three laws of responsible robotics," *IEEE Intelligent Systems*, vol. 24, pp. 14–20, July 2009.

[2] A. Polyportis and N. Pahos, "Navigating the perils of artificial intelligence: a focused review on chatgpt and responsible research and innovation," *Humanities and Social Sciences Communications*, vol. 11, p. 107, Jan. 2024.

[3] B. C. Stahl and D. Eke, "The ethics of chatgpt – exploring the ethical issues of an emerging technology," *International Journal of Information Management*, vol. 74, p. 102700, Feb. 2024.

[4] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, pp. 389–399, Sept. 2019.

[5] Z. Tóth, R. Caruana, T. Gruber, and C. Loebbecke, "The dawn of the ai robots: Towards a new framework of ai robot accountability," *Journal of Business Ethics*, vol. 178, pp. 895–916, July 2022.

[6] S. Neupane, S. Mitra, I. A. Fernandez, S. Saha, S. Mittal, J. Chen, N. Pillai, and S. Rahimi, "Security considerations in ai-robotics: A survey of current methods, challenges, and opportunities," *IEEE Access*, vol. 12, pp. 22072–22097, Jan. 2024.

[7] Z. Alsulaimawi, "A privacy filter framework for internet of robotic things applications," in *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 262–267, IEEE, May 2020.

[8] J.-P. A. Yaacoub, H. N. Noura, O. Salman, and A. Chehab, "Robotics cyber security: vulnerabilities, attacks, countermeasures, and recommendations," *International Journal of Information Security*, vol. 21, pp. 115–158, Mar. 2021.

[9] L. Levinson, C. Nippert-Eng, R. Gomez, and S. Sabanović, "Snitches get unplugged: Adolescents' privacy concerns about robots in the home are relationally situated," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 423–432, ACM, Mar. 2024.

[10] A. A. Biswas, M. S. Zulfiker, M. M. Rahman, M. R. Jani, and M. M. Anwar, "Data privacy and security analysis for mental health chatbot applications," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 1245–1249, IEEE, Mar. 2025.

[11] S. Ramya, M. Radhika, S. Roychoudri, S. Senthil, G. Arunachalam, and M. Muthulekshmi, "Secure and responsive human-robot interaction for dementia patients using lstm and iot," in *2024 3rd International Conference for Advancement in Technology (ICONAT)*, pp. 1–6, IEEE, Sept. 2024.

[12] I. G. I. on Ethics of Autonomous and I. Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. IEEE Standards Association, 2019.

[13] B. Irfan, S. Kuoppamäki, and G. Skantze, "Recommendations for designing conversational companion robots with older adults through foundation models," *Frontiers in Robotics and AI*, vol. 11, May 2024.

[14] F. Martín, E. Soriano-Salvador, J. M. Guerrero, G. G. Múzquiz, J. C. Manzanares, and F. J. Rodríguez, "Towards a robotic intrusion prevention system: Combining security and safety in cognitive social robots," *Robotics and Autonomous Systems*, p. 104959, Mar. 2025.

[15] W. Salhab, D. Ameyed, F. Jaafar, and H. Mcheick, "A systematic literature review on ai safety: identifying trends, challenges, and future directions," *IEEE Access*, vol. 12, pp. 131762–131784, Jan. 2024.

[16] B. Hutler, T. N. Rieder, D. J. H. Mathews, D. A. Handelman, and A. M. Greenberg, "Designing robots that do no harm: understanding the challenges of ethics for robots," *AI and Ethics*, vol. 4, pp. 463–471, May 2024.

[17] T. Reinmund, P. Salvini, L. Kunze, M. Jirotka, and A. F. T. Winfield, "Variable autonomy through responsible robotics: Design guidelines and research agenda," *ACM Transactions on Human-Robot Interaction*, vol. 13, pp. 1–36, Mar. 2024.

[18] M. Charikar, J. Steinhardt, and G. Valiant, "Learning from untrusted data." arXiv preprint arXiv:1611.02315, Jan. 2016.