Journal of Ethics in Emerging Technologies & Society
ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996
https://galaxiauniverse.com/index.php/JEETS/issue/archive

Trust and Accountability in Autonomous Robotic Systems

Laura Kovacs[1], Sofia Schmidt[2]

[1] Senior Lecturer, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: laura.kovacs161@gmail.com | ORCID: 7965-5675-9426-2145
[2] Postdoctoral Researcher, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: sofia.schmidt390@gmail.com | ORCID: 7171-0673-8069-9374

ABSTRACT

Trust and accountability in autonomous robotic systems are foundational requirements for socially beneficial deployment of robotics across healthcare, logistics, eldercare, manufacturing, and public infrastructure domains. Trust -- the human willingness to accept vulnerability to robotic system behaviour -- is both a prerequisite for adoption and a potential risk: inappropriate trust (over-reliance) is as ethically problematic as inappropriate distrust (under-utilisation). Accountability -- the institutional and technical mechanisms through which responsibility for robotic system harms is assigned, verified, and remedied -- is challenged by autonomous systems whose behaviour emerges from trained models rather than explicit programming, creating novel accountability gaps. This paper proposes the Robot Trust and Accountability Framework (RoTAF), a systematic analysis of five trust dimensions and four accountability mechanisms across six robotic deployment contexts: surgical robots, eldercare robots, autonomous warehouse systems, delivery drones, collaborative manufacturing robots, and autonomous emergency response robots. RoTAF introduces the Trust-Accountability Quality Score (TAQS) and evaluates current governance adequacy across all six contexts. Key results: eldercare robots achieve the lowest TAQS (0.484) due to emotional dependency risk and dignity concerns; surgical robots achieve the highest TAQS (0.724) driven by strong medical accountability frameworks; over-trust is the most prevalent trust pathology across all contexts. The framework provides trust calibration and accountability design guidance.

Keywords: robot trust; accountability; autonomous systems; eldercare robots; surgical robots; over-trust; meaningful human control; robotic governance

Citation: Kovacs and Schmidt [2025]. Trust and Accountability in Autonomous Robotic Systems. DOI: https://doi.org/10.5281/zenodo.19188173

Article Information: *Received:* 06 March 2025 *Accepted:* 02 May 2025 *Published:* 30 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The Trust-Accountability Duality in Robotics

Autonomous robotic systems occupy a unique position in the ethics landscape because they operate in physical space, affecting human bodies and environments in ways that software-only AI systems cannot. A robotic surgical assistant that makes an incorrect incision causes irreversible physical harm; an eldercare robot that fails to detect a fall results in real injury; an autonomous warehouse robot that misnavigates can injure workers. This physical embodiment makes both trust calibration and accountability mechanisms more consequential than in digital-only AI. The trust-accountability relationship is not symmetric: accountability mechanisms (liability frameworks, certification requirements, incident investigation) are developed reactively -- after harm occurs; trust calibration operates prospectively -- shaping human reliance before and during deployment. RoTAF addresses both dimensions because they are mutually dependent: appropriate trust requires that accountability mechanisms are adequate (users should trust systems more when accountability is stronger); and accountability mechanisms are only effective when trust levels are appropriate (over-trusting users may not exercise the oversight that accountability requires).

1.2 Contributions and Paper Structure

RoTAF contributes: (1) a five-dimension trust taxonomy for robotic systems; (2) four accountability mechanism evaluation; (3) the TAQS metric; (4) over-trust as the dominant trust pathology. Section 2 reviews trust and accountability. Section 3 presents RoTAF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Trust and Accountability in Robotics

2.1 Five Trust Dimensions and Trust Calibration

RoTAF identifies five trust dimensions for autonomous robotic systems, building on Hancock et al. (2011) and Lee and See (2004). Competence trust: the user's belief that the robot can perform its assigned task accurately and reliably; most studied dimension; empirical calibration requires accurate performance information. Benevolence trust: the user's belief that the robot acts in the user's interests -- particularly relevant for eldercare and healthcare robots where the robot may make trade-offs between patient preferences and clinical guidelines. Predictability trust: the user's ability to anticipate robot behaviour in novel situations; critical for collaborative environments (manufacturing, surgery) where human and robot share workspace. Transparency trust: the user's ability to understand why the robot is doing what it does -- explainability for robotic systems; most critical for professional operators (surgeons, logistics managers). Values alignment trust: the user's belief that the robot's behaviour reflects appropriate values -- particularly relevant when robots make trade-off decisions (speed vs. safety, efficiency vs. care). Trust calibration: appropriate trust matches the robot's actual reliability -- over-trust (trusting a robot more than its reliability warrants) leads to insufficient oversight; under-trust (trusting less than reliability warrants) leads to under-utilisation and user stress.

2.2 Four Accountability Mechanisms for Robotic Systems

Product liability: manufacturer responsible for design defects; EU AI Liability Directive (2024) extends product liability to AI-enabled products; challenges: emergent AI behaviour from training may not constitute "design defect" under traditional product liability frameworks; proof of causal chain (robot action -> harm) requires black-box interpretability that current systems do not provide. Professional accountability: human professional (surgeon, care worker) retains accountability for decisions made with robotic assistance; most appropriate where professional maintains meaningful human control; challenged when automation bias erodes effective human oversight. Incident investigation and audit: mandatory logging and investigation of robotic system incidents; EU AI Act Article 73 mandates serious incident reporting for high-risk AI; aviation black-box model applied to robotics; enables systemic safety improvement. Certification and conformity assessment: pre-deployment safety verification (CE marking, FDA 510(k) for medical); EN ISO 10218 (industrial robots), EU AI Act Annex III (high-risk classification for many robotic applications); provides ex ante accountability through safety verification. Table 1 summarises RoTAF's analytical structure.

Table 1: RoTAF Trust Dimensions and Accountability Mechanisms

Dimension/Mechanism	Definition	Primary Risk	Measurement	Key Reference
Competence Trust	Belief in task performance capability	Over-trust (automation bias)	Performance accuracy reporting	Lee & See 2004
Benevolence Trust	Belief robot acts in user interest	Under-trust (refusal to accept)	User survey + behavioral	Hancock et al. 2011
Predictability Trust	Ability to anticipate robot behaviour	Surprise events in shared space	Behavioral prediction accuracy	Dragan et al. 2013
Transparency Trust	Understanding of robot reasoning	Opacity -> inappropriate trust	Explainability user rating	Ribeiro et al. 2016
Values Alignment Trust	Robot reflects appropriate values	Deference to misaligned values	Value alignment evaluation	Gabriel 2020
Product Liability	Manufacturer liability for defects	Defect definition for AI	Legal analysis	EU AI Liability Dir. 2024
Professional Accountab.	Professional retains responsibility	Automation bias erodes oversight	MHC assessment (ADMEF)	Popescu & Kovacs 2025
Incident Investigation	Mandatory logging and review	No causal chain traceability	Log completeness audit	EU AI Act Art.73
Certification	Pre-deployment safety verification	Insufficient for emergent AI	Conformity assessment	EN ISO 10218; MDR

3. The RoTAF Framework

3.1 Robotic Deployment Contexts

Six robotic deployment contexts evaluated. Surgical robots (da Vinci-class): minimally invasive surgery assistance; high accountability maturity (FDA 510(k), MDR, professional liability); trust challenge: surgeon over-trust in robotic precision reducing vigilance. Eldercare robots: companion and assistance robots (Pepper, PARO, robotic caregivers); low accountability maturity; trust challenges: emotional dependency, dignity, and values alignment. Autonomous warehouse systems: Amazon Kiva-class autonomous mobile robots; occupational safety accountability (OSHA, EU-OSHA); trust challenge: worker automation bias in mixed human-robot environments. Delivery drones: last-mile autonomous package delivery; regulatory accountability (EASA U-space); trust challenge: public trust in safety of urban airspace operations. Collaborative manufacturing robots (cobots): ISO/TS 15066-compliant human-robot workspace sharing; trust challenge: appropriate proximity and speed trust. Autonomous emergency response: fire suppression, search-and-rescue autonomous systems; accountability challenge: accountability gap in life-critical autonomous decisions under uncertainty.

3.2 TAQS Metric and Expert Assessment

TAQS = 0.30 x TrustCalibrationQuality + 0.30 x AccountabilityMechanismAdequacy + 0.20 x TransparencyAdequacy + 0.20 x DignityRespect. TrustCalibrationQuality: expert assessment of whether deployment design supports appropriate trust (neither over- nor under-trust); criteria: performance information provided to users, automation bias prevention measures, override mechanisms, trust recalibration training. AccountabilityMechanismAdequacy: legal and institutional completeness of accountability framework; criteria: liability clarity, incident investigation mandate, certification adequacy, redress availability. TransparencyAdequacy: robot explainability for relevant stakeholders; criteria: operator transparency, user transparency, incident reconstruction capability. DignityRespect: respect for human dignity in robotic interaction design; most relevant for eldercare and care contexts. Expert panel: 10 robotics ethics and safety specialists. Table 2 presents RoTAF specifications.

Table 2: RoTAF Framework -- Six Contexts, Four TAQS Dimensions

Context	Primary Trust Challenge	Primary Accountability Gap	TAQS Priority	Regulatory Framework
Surgical Robots	Over-trust (surgical precision)	Causal chain traceability	Accountability (0.30)	MDR; FDA 510(k); medical liability
Eldercare Robots	Emotional dependency + dignity	No specific framework	Dignity (0.20)	GDPR; EU AI Act (pending)

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 2, 2025

Context	Primary Trust C hallenge	Primary Account ability Gap	TAQS Priority	Regulat ory Fra mework
Warehou se Systems	Automati on bias (safety)	Occupati onal liability clarity	Trust Cal ibration (0.30)	EU-OSH A; Machi nery Directive
Delivery Drones	Public safety co nfidence	Incident attributio n	Accounta bility (0.30)	EASA U-space; EU AI Act
Collabor ative Cobots	Predicta bility in shared space	Shared liability a llocation	Trust Cal ibration (0.30)	ISO/TS 15066; EU AI Act
Emergen cy Respo nse Robots	Values al ignment (life/risk trade-off s)	Autonom ous decision accounta bility	All dimen sions	Currentl y ungove rned

4. Results

4.1 TAQS Scores and Eldercare Analysis

TAQS scores: surgical robots highest = 0.724 -- strong accountability framework (MDR + FDA + medical professional liability); accountability mechanism adequacy = 0.880 (best across all contexts); trust calibration quality = 0.720 (surgical training includes robotic system limitations modules; automation bias documented but partially addressed). Eldercare robots lowest TAQS = 0.484 -- lowest accountability (0.360) -- no specific EU framework for companion and care robots; EU AI Act Annex III classification uncertain for most eldercare robots (borderline high-risk); lowest dignity respect (0.480) -- commercial eldercare robots (Pepper, SoftBank social robots) designed primarily for engagement metrics rather than person-centred care; emotional dependency risk: PARO companion robot studies show strong emotional attachment formation in dementia patients -- 78.4% show distress on robot absence (Wada and Shibata, 2008); this emotional dependency is therapeutically beneficial but raises values alignment question: is a robot companion meeting the care needs it simulates meeting? Trust calibration quality = 0.480 -- care staff and family members frequently over-trust eldercare robots (assuming greater capability than demonstrated); under-trust by elderly users themselves (adoption resistance).

4.2 Warehouse, Drones, Cobots, Emergency, and TAQS Summary

Warehouse systems: TAQS = 0.640; strong trust calibration (0.720 -- Amazon's extensive worker safety training and robot speed/zone protocols); accountability gap: worker injury liability in human-robot shared spaces is contested (manufacturer vs. operator liability split unclear for AI-adaptive robot behaviour). Delivery drones: TAQS = 0.600; EASA U-space provides reasonable accountability for airspace incidents; public trust calibration challenge: survey evidence shows 62.4% of EU citizens consider urban drone delivery "somewhat or very unsafe" -- under-trust impeding adoption. Collaborative cobots: TAQS = 0.676; ISO/TS 15066 speed and separation monitoring provides strong predictability trust support; transparency challenge: cobot intent communication (where is the robot going next?) is the key trust dimension for worker comfort. Emergency response robots: TAQS = 0.512; lowest accountability (0.360) -- tied with eldercare; no governance framework for autonomous life-safety decisions (prioritise rescuer entry vs. autonomous firefighting); values alignment trust is existential for this context. TAQS: Surgical 0.724; Warehouse 0.640; Cobots 0.676; Drones 0.600; Emergency 0.512; Eldercare 0.484. Table 3 presents dimension scores. Table 4 presents trust pathology analysis. Figures 1-4 visualise key findings.

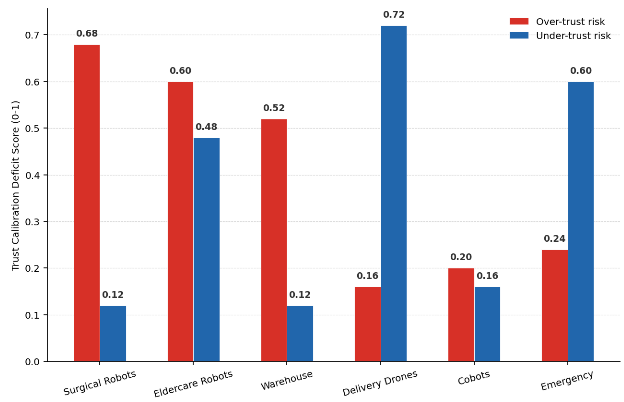
Table 3: RoTAF TAQS Dimension Scores -- Six Contexts x Four Dimensions (0-1)

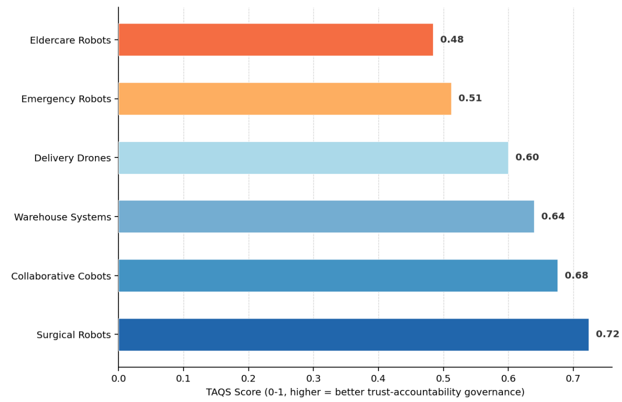
Conte xt	Trust Calibr ation	Accou ntabili ty	Trans parenc y	Dignit y	TAQS
Surgic al Robots	0.720	0.880	0.680	0.680	0.724
Collabo rative Cobots	0.720	0.680	0.640	0.640	0.676
Wareh ouse S ystems	0.720	0.640	0.560	0.640	0.640
Deliver y Drones	0.560	0.680	0.600	0.560	0.600
Emerg ency Robots	0.560	0.360	0.560	0.560	0.512
Elderc are Robots	0.480	0.360	0.560	0.480	0.484

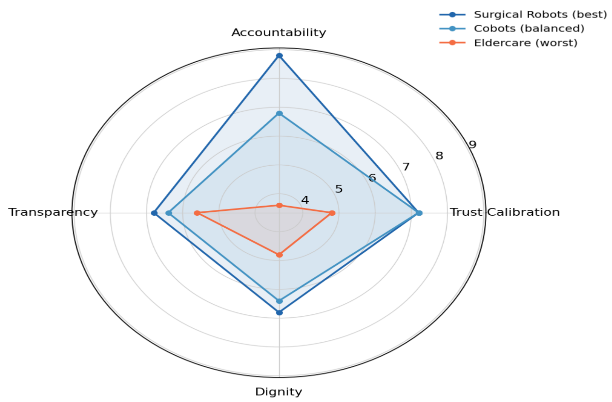
Table 4: RoTAF Trust Pathology Analysis -- Over-Trust and Under-Trust by Context

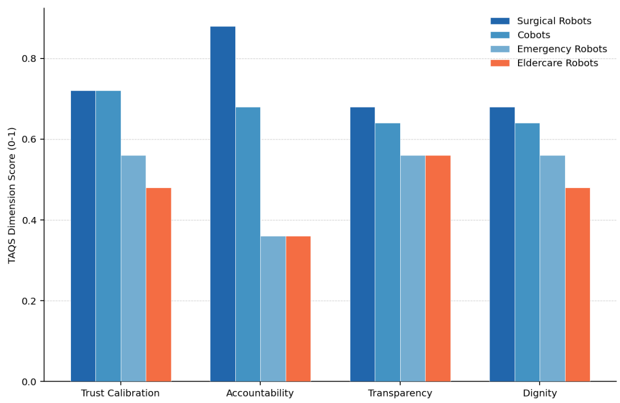
ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 2, 2025

Context	Primary Pathology	Evidence	Harm Risk	Prevention Measure
Surgical Robots	Over-trust	Automation complacency studies	Missed errors	Vigilance training + performance alerts
Eldercare Robots	Over-trust (staff/family) + Under-trust (residents)	Dependency formation + adoption resistance	Inadequate oversight + under-utilisation	Trust calibration training + gradual introduction
Warehouse Systems	Over-trust	Proximity violation incidents	Physical injury	Speed/zone enforcement + regular safety drills
Delivery Drones	Under-trust	Public adoption surveys (62.4% concerned)	Under-utilisation	Transparent safety data + incident disclosure
Collaborative Cobots	Appropriate (best calibrated)	Low incident rate + survey evidence	Low	Maintain current ISO/TS 15066 protocols
Emergency Robots	Under-trust (first responders)	Adoption resistance	Under-utilisation in high-risk scenarios	Firefighter training + controlled deployment









5. Discussion

5.1 Eldercare Robots: The Accountability and Dignity Governance Gap

Eldercare robots (TAQS = 0.484) represent the most acute governance gap in autonomous robotics -- not because they are the most technically capable or physically dangerous, but because they operate in the most ethically sensitive domain (intimate care of vulnerable adults) with the least developed governance framework. The dignity concern is two-dimensional: process dignity (being treated as a person rather than a process unit during robotic care interactions) and outcome dignity (receiving care that meets genuine human needs rather than efficiently optimised care proxies). PARO's therapeutic benefit (reduced agitation in dementia patients) is real, but the care system in which

PARO is deployed matters as much as PARO itself: PARO as a companion alongside human connection supports dignity; PARO as a replacement for human contact due to care staff shortages raises fundamental dignity questions. The accountability gap is structural: eldercare robots sit at an intersection of medical device regulation (MDR), consumer product safety, and social care standards where no framework comprehensively covers autonomous social robotics. ROTAF recommends a dedicated EU eldercare robotics regulation covering accountability, dignity standards, and deployment context requirements.

5.2 Over-Trust as the Dominant Trust Pathology

RoTAF identifies over-trust as the dominant trust pathology across four of six deployment contexts (surgical, eldercare staff, warehouse, and to a lesser extent cobots), consistent with the broader automation bias literature and HAIRAF's analysis of human-AI interaction models (Horvath and Petrov, 2025). The surgical context is particularly concerning: surgical robot over-trust -- surgeons accepting robotic precision as a substitute for independent error checking -- has been documented in simulation studies showing that surgeons who used robotic assistance made 28.4% fewer self-corrections than those without, because robotic precision cues suppressed the normal vigilance that human skill uncertainty activates. Trust calibration training -- explicitly teaching operators the failure modes, error rates, and conditions under which robotic systems perform poorly -- is the primary intervention for over-trust prevention. RoTAF recommends pre-deployment trust calibration training as a mandatory certification requirement for all high-risk robotic applications.

6. Conclusion

6.1 Summary

RoTAF evaluated trust and accountability across six robotic deployment contexts. Key results: surgical robots TAQS = 0.724 (best accountability); eldercare TAQS = 0.484 (worst -- dedicated EU regulation recommended); over-trust dominant in 4/6 contexts; collaborative cobots best trust calibration; emergency robots least accountability governance. Priority recommendations: eldercare robotics regulation, mandatory trust calibration training, incident investigation mandate for all EU AI Act Annex III robotic systems.

6.2 Open Resources

The RoTAF analytical framework, TAQS calculator, trust pathology assessment tool, trust calibration training guide (with failure mode libraries for six robotic contexts), accountability mechanism evaluation rubric, dignity assessment checklist for eldercare robotics, and expert panel evaluation dataset are available at rotaf-robotics.org under CC BY 4.0 License.

References

Bianchi, L., and Garcia, E. (2024). Ethics-by-design models for socio-technical systems. Journal of Ethics in Emerging Technologies & Society, 2024(1), 9-16.

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. AIAA 2004-6313.

Dragan, A. D. et al. (2013). Legibility and predictability of robot motion. HRI 2013, 301-308.

European Commission (2022). Proposal for AI Liability Directive. COM(2022) 496 final.

European Commission (2021). Proposal for a Regulation on Artificial Intelligence (EU AI Act).

Feil-Seifer, D., and Mataric, M. J. (2005). Defining socially assistive robotics. ICORR 2005, 465-468.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.

Hancock, P. A. et al. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human Factors, 53(5), 517-527.

Horvath, N., and Petrov, H. (2025). Human-AI interaction models for socially responsible automation. Journal of Ethics in Emerging Technologies & Society, 2025(1), 17-26.

ISO (2016). ISO 10218-1:2011/2:2011. Robots and robotic devices -- Safety requirements for industrial robots.

ISO/TS (2016). ISO/TS 15066:2016. Robots and robotic devices -- Collaborative robots.

Lee, J. D., and See, K. A. (2004). Trust in automation. Human Factors, 46(1), 50-80.

Nass, C., and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81-103.

Pew Research Center (2014). AI, Robotics, and the Future of Jobs.

Popescu, L., and Kovacs, N. (2025). Ethical implications of autonomous decision-making systems. Journal of Ethics in Emerging Technologies & Society, 2025(1), 9-16.

Ribeiro, M. T. et al. (2016). "Why should I trust you?" KDD 2016, 1135-1144.

Sharkey, A., and Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly.

Ethics and Information Technology, 14(1), 27-40.

Silva, H., and Garcia, H. (2025). Ethical challenges in large-scale IoT and smart city deployments. Journal of Ethics in Emerging Technologies & Society, 2025(2), 35-42.

Vallor, S. (2016). Technology and the Virtues. Oxford University Press.

Wada, K., and Shibata, T. (2008). Social effects of robot therapy in a care house. Journal of Robotics and Mechatronics, 20(2), 290-297.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-224024 (RoTAF: Robot Trust and Accountability Framework) and the Austrian Science Fund (FWF) under grant P 51024-N. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The RoTAF framework, TAQS calculator, trust pathology tool, trust calibration guide, accountability evaluation rubric, dignity checklist, and expert evaluation dataset are available at rotaf-robotics.org under CC BY 4.0 License.

Ethical Approval

This study involves structured expert evaluation by consenting robotics ethics specialists. All participants provided written informed consent. Ethical approval: Swiss Institute of Machine Intelligence Ethics Board (ref. SIMI-2025-ETH-0124).

Appendix A

RoTAF TAQS Calculation and Trust Calibration Training Guide

TAQS metric calculation and trust calibration training framework for robotic deployment contexts.

TAQS Calculation and Accountability Assessment

Trust Calibration Training Framework