information

MDPI

Systematic Review

Cybersecurity and Privacy for Co-Creative Robotics: Protecting Trust Without Constraining Creative Autonomy †

Eda Marchetti [1,*], Sanaz Nikghadam-Hojjati [2,*], Antonello Calabrò [1] and José Barata [2]

[1] Istituto di Scienza e Tecnologie dell'Informazione (ISTI), Consiglio Nazionale Delle Ricerche (CNR), 56124 Pisa, Italy; antonello.calabro@isti.cnr.it
[2] UNINOVA-CTS, LASI, NOVA School of Science and Technology, Nova University Lisbon, NOVA-FCT Campus, 2829-516 Caparica, Portugal; jab@uninova.it
* Correspondence: eda.marchetti@isti.cnr.it (E.M.); sanaznik@uninova.pt (S.N.-H.)
† Special Issue: Emerging Research in Computational Creativity and Creative Robotics.

Abstract

Co-Creative Robotics combines computational creativity, robotic embodiment, and human–robot collaboration to support or generate creative behavior in physical and social environments. As these systems become more autonomous, data-intensive, and interactive, cybersecurity and privacy can no longer be treated as external safeguards added after creative functionality has been designed. This PRISMA-informed review investigates whether principles of cybersecurity-by-design and privacy-by-design can be integrated into Co-Creative Robotics without constraining creativity, autonomy, and user agency. The database search covered ACM Digital Library, Google Scholar, IEEE Xplore, Scopus, SpringerLink, and Web of Science, and was complemented by two focused backward and forward snowballing iterations. From 623 database records, the final synthesis includes 27 primary studies. The results show that direct literature combining cybersecurity, privacy, and Co-Creative Robotics remains limited, but evidence from creative HRI, social-robot privacy, cyber-physical security, privacy-preserving interaction design, security modeling, and robot ethics supports a conditional answer. Integration is feasible when security and privacy mechanisms are adaptive, explainable, participatory, context-sensitive, and lifecycle-aware. However, the evidence on transparency-oriented privacy mechanisms is mixed: improvements in awareness or acceptance do not consistently translate into reduced disclosure or greater perceived safety. Rigid controls may constrain creative exploration, whereas well-designed controls can support trust, accountable autonomy, safe embodiment, privacy-aware interaction, provenance, and agency-preserving creativity. The review proposes a conceptual lifecycle-oriented research agenda for secure and privacy-aware Co-Creative Robotics.

Keywords: Co-Creative Robotics; creative robotics; computational creativity; cybersecurity-by-design; privacy-by-design; social robots; human–robot interaction; security modeling; privacy-preserving sensing; robot autonomy

Check for updates

Academic Editor: Aristeidis Karras

Received: 13 July 2026
Revised: 5 August 2026
Accepted: 7 August 2026
Published: 11 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Co-Creative Robotics (CCR) [1] emerges at the intersection of computational creativity, robotic embodiment and collaborative human–machine interaction. Unlike conventional robots designed primarily for predefined tasks, co-creative robots are expected to support or enact creative behavior, contribute to open-ended tasks, adapt to changing contexts, and interact with humans in ways that preserve agency, initiative, and expressive freedom [1].

This makes CCR particularly relevant to the agenda of computational creativity and creative robotics, where artificial and robotic systems are not only tools for automation but also partners, scaffolds, or agents in creative processes.

The creative dimension of robotics is not merely a matter of generating novel outputs. It also concerns how users experience collaboration, how autonomy is negotiated, how the robot makes suggestions, how it responds to ambiguity, and how creative authority is distributed between the human and the machine. Empirical evidence from social robots used as creativity-eliciting agents shows that embodied agents can stimulate verbal, figural, and constructional creativity through demonstration and scaffolding behaviors [2]. Likewise, studies of artificial intelligence (AI) assistance in engineering design show that AI can change the exploration strategy, novelty, and effectiveness of human design work, especially when task complexity increases [3]. These results suggest that creative autonomy is an interactional property: it depends on the way technical mechanisms are embedded in the task, interface, and collaboration structure.

At the same time, the same characteristics that make CCR promising also expand its risk surface. Co-creative robots may collect multimodal data, process images, speech, gestures, gaze, affective cues, and contextual signals, learn from user behavior, generate or modify creative artifacts, operate in shared physical spaces, and adapt their actions over time. In the European legal context, data protection by design and by default is explicitly established in Article 25 of the General Data Protection Regulation (GDPR), making privacy a design requirement rather than an optional add-on [4]. From the cybersecurity side, security-by-design has become a broad response to the limitations of reactive protection, especially in digital systems where convenience, interconnection, and user-centered design can conflict with robust protection [5].

This creates a central tension for CCR. If cybersecurity and privacy controls are implemented as rigid restrictions, they may reduce spontaneity, exploration, co-adaptation, and the sense of agency that are essential to creative collaboration. If they are ignored or postponed, the resulting system may undermine trust, expose private data, enable manipulation, weaken accountability, or produce unsafe physical behavior. Therefore, the research problem addressed in this review is not whether cybersecurity and privacy are relevant to CCR, but how to adapt them to preserve the qualities that make co-creative interaction valuable.

Existing work remains fragmented. Studies on Co-Creative Robotics and creative human–robot interaction primarily focus on creativity, collaboration, autonomy, social interaction, and user experience rather than on cybersecurity or privacy as first-order design concerns [1–3]. In contrast, robotics cybersecurity studies tend to address anomaly detection, cyber-physical attacks, testbeds, threat modeling, digital-twin verification, or secure architectures, often with limited attention to creativity, agency, or co-creative interaction [6–9]. Privacy-oriented and ethical studies provide important insights into agency, control, accountability, trust, disclosure, and social justice, but these insights are not always translated into concrete lifecycle controls for secure and privacy-aware CCR [10–28]. Therefore, a review is necessary to connect these strands and derive design implications for secure and privacy-aware co-creative robotics. Against this background, the review addresses the following main research question:

Main RQ: Is it possible to integrate principles of cybersecurity and privacy into Co-Creative Robotics without constraining creativity and autonomy?

To operationalize this overarching question, the review is organized around five sub-research questions:

SRQ1: What are the primary cybersecurity and privacy challenges, issues, and opportunities?

SRQ2: Which current and emerging application domains are being addressed?

SRQ3: What current and emerging methods are being proposed or are transferable?

SRQ4: What technical, ethical, and regulatory challenges affect integration?

SRQ5: What future directions emerge throughout the lifecycle, from data collection and creative model training to robotic embodiment and interaction?

The contribution of this paper is fourfold. First, it provides a PRISMA-informed selection and synthesis of studies relevant to cybersecurity, privacy, and Co-Creative Robotics. Second, it distinguishes core creative-robotics evidence from related robotics and HRI studies, social-robot privacy studies, and transferable cybersecurity and privacy methods. Third, it answers the main research question and the five sub-research questions through a thematic synthesis of the final corpus. Fourth, it proposes a lifecycle-oriented research agenda that reframes cybersecurity-by-design and privacy-by-design as enablers of trust, accountable autonomy, and agency-preserving creativity, rather than as external constraints imposed on the creative process.

The remainder of the paper is organized as follows. Section 2 introduces the conceptual background and related work on Co-Creative Robotics, cybersecurity-by-design, privacy-by-design, creative autonomy, trust, and lifecycle risks. Section 3 presents the PRISMA-informed methodology, including the search strategy, eligibility criteria, screening process, full-text assessment, two-iteration snowballing, quality assessment, and data extraction procedure. Section 4 reports the composition and quality of the final corpus and presents the thematic synthesis by research question. Section 5 discusses the cross-cutting implications of the findings, as well as limitations and future research directions. Section 6 concludes the paper.

2. Background and Related Work

This section clarifies the conceptual and methodological foundations needed to interpret the review results, focusing on the specific properties of CCR that make cybersecurity-by-design and privacy-by-design different from their application in conventional software or non-embodied AI systems.

2.1. Conceptual Foundations of Co-Creation and Creative Autonomy

CCR is treated here as an embodied form of computational creativity in which the robotic system participates in a shared creative process rather than merely executing a predefined task [1]. Its theoretical roots are connected to computational creativity, which studies how artificial systems can model, support, evaluate, or generate creative behavior [29,30], to creativity support tools, which emphasize technologies that make users more creative more often rather than simply automating the final output [31], and to recent work on human–robot collaborative creativity, which highlights the cognitive, affective, social, creative, and ethical dimensions of robot-supported creative collaboration [32].

This distinction matters because CCR adds embodiment, situated interaction, and physical or social presence to the creative process. In a disembodied creative AI system, protection can often be framed around data, models, prompts, and generated artifacts. In CCR, the same issues are coupled with sensing, motion, proxemic behavior, turn-taking, gaze, manipulation, and the user's perception of being observed, assisted, interrupted, or influenced by an autonomous agent. Therefore, security and privacy decisions affect not only information flows but also the quality of creative participation.

Creative autonomy is used in this review to denote negotiated initiative rather than unconstrained robot action. It includes users' ability to explore alternatives, reinterpret robotic suggestions, reject or modify outputs, understand when and why the robot intervenes, and retain meaningful influence over the creative process. The empirical studies

included in the corpus illustrate two complementary dimensions of this concept: embodied scaffolding can stimulate creativity in social-robot interaction [2], while AI assistance can reshape search behavior, novelty, and effectiveness in complex engineering design tasks [3]. These findings support the view that autonomy in CCR is relational and socio-technical: it emerges from the interaction among task structure, interface design, robot behavior, and user control.

2.2. Cybersecurity-by-Design and Privacy-by-Design

Cybersecurity-by-design refers to the integration of security requirements, risk reasoning, and protective controls from the earliest stages of system development. In cyber-physical and manufacturing contexts, this principle has also been linked to design-time architectural choices that embed protection across hardware and software layers [33]. It contrasts with reactive security approaches in which protection is added only after architecture, interaction flows, and deployment assumptions have already been fixed. Systems security engineering provides a useful foundation for this view because it treats security as a system property to be addressed across requirements, architecture, implementation, verification, operation, maintenance, and disposal [34], while lifecycle-oriented infrastructures such as ACCURATE extend this logic toward continuous post-development assessment and monitoring [35]. This lifecycle perspective is especially important for robotics, where software, sensors, actuators, networks, users, and operating environments interact continuously.

In CCR, security-by-design must account for threats to data, models, communication channels, sensor integrity, actuation, runtime adaptation, and creative outputs. Work on digital security by design highlights the broader shift from reactive protection to design-time security reasoning in interconnected digital ecosystems [5]. Robotics-specific evidence reinforces this need: anomaly detection for robotic arms shows that cyber-physical deviations may not be visible from network data alone [6], while cybersecurity test ranges for autonomous shuttles demonstrate the value of controlled environments for validating attacks and mitigations before real-world deployment [7].

For CCR, privacy-by-design must be interpreted as an interactional and creative requirement, not only as a data-protection principle. Building on the general privacy-by-design premise that privacy should be embedded in systems and default settings [36], the relevant CCR privacy problem is not limited to whether personal data are collected lawfully; it also concerns how sensing, memory, adaptation, and creative generation reshape the user's control over ideas, drafts, gestures, preferences, emotions, and co-created outputs. In this sense, privacy protection is closely connected to creative agency: users should be able to understand what the robot observes, decide what may be retained or reused, and contest or redirect the transformation of personal and creative traces during the interaction.

The implication is that cybersecurity-by-design and privacy-by-design cannot be reduced to restrictive access rules. In CCR, they should help define what can be sensed, what should remain local, what users can inspect or contest, how creative data are retained or deleted, how generated artifacts are attributed, and how robot behavior remains safe and accountable under uncertainty. Protection becomes compatible with creativity when it is proportionate, understandable, adjustable, and aligned with the user's creative agency. Recent work on co-creative cobot communities similarly highlights trust as a mediating condition for creative cooperation in mixed human–robot settings [37].

2.3. Trustworthy and Responsible CCR Across the Lifecycle

Trustworthy and responsible AI frameworks provide a bridge between technical protection and human-centered robotics. The European trustworthy AI guidelines connect technical robustness with human agency, privacy and data governance, transparency, fair-

ness, societal well-being, and accountability [38]. IEEE Ethically Aligned Design similarly frames autonomous and intelligent systems in terms of human well-being, values, accountability, transparency, and awareness of misuse [39]. These perspectives are relevant to CCR because co-creative robots may influence not only task execution, but also ideation, authorship, expression, collaboration, and the user's sense of creative ownership.

A lifecycle view is necessary because the risks of CCR change as the system moves from conception to operation. During data collection, the main issues concern minimization, informed consent, bystander privacy, visual privacy, and the visibility of sensing. During training or adaptation, risks include data poisoning, model leakage, unauthorized reuse of creative datasets, and loss of control over model-derived outputs. During interaction, risks include manipulation, over-reliance, unsafe suggestions, adversarial prompts, loss of initiative, and physical consequences of cyber compromise. During deployment and governance, risks concern insecure updates, insufficient monitoring, weak documentation, unclear provenance, and accountability gaps.

Threat modeling can help translate these risks into design decisions. General threat modeling supports the early identification of assets, trust boundaries, attack paths, and mitigations [40]. Privacy threat modeling extends this reasoning to privacy-specific threats such as linkability, identifiability, information disclosure, unawareness, and non-compliance [41]. Socio-technical security modeling is particularly relevant for CCR because the definition of what needs protection depends on the perspective of users, designers, operators, and affected stakeholders. Explicitly, the modeler, orientation, referent object, and stance of a security model are also important [42].

2.4. Related Work Landscape

The studies considered in this review are not treated as a homogeneous body of evidence. They are organized according to the analytical role they play in answering the research questions. This distinction is necessary because the intersection between CCR, cybersecurity, and privacy is still emerging, and relevant evidence is distributed across creative robotics, social-robot privacy, cyber-physical security, security engineering, and responsible robotics. Accordingly, the relevant literature can be organized into four complementary strands.

First, studies on CCR, creative HRI, computational creativity, and creativity-support technologies define the qualities that protection mechanisms should preserve, including exploration, scaffolding, novelty, initiative, interpretive flexibility, and meaningful human participation [1–3,29–31]. Second, studies on social robots, domestic robots, healthcare robots, privacy settings, warnings, consent, disclosure, sensor-state communication, and privacy-aware motion show how privacy and agency are experienced and negotiated in embodied interaction [11,15–28].

Third, work on robotic anomaly detection, cyber ranges, digital-twin security, threat modeling, security modeling, medical-device security, and agentic-AI security provides transferable methods for identifying, testing, monitoring, and governing risks in autonomous or cyber-physical systems [6–9,40,42,43]. Fourth, studies on human–robot collaboration, embodied AI ethics, autonomous cyber-physical systems, care robotics, LLM-enabled multi-robot systems, technology abuse, blockchain provenance, decentralized multi-robot coordination, and federated learning provide ethical and governance concepts such as accountability, contextual autonomy, inclusion, ownership, misuse, and responsible governance [10,12,13,44–49].

This organization clarifies the background landscape of the review while avoiding the assumption that all sources are equivalent evidence. The detailed study-level classification

of the four evidence groups, including their different levels of proximity to CCR and their role in the thematic synthesis, is provided in Section 4.1.

3. Materials and Methods

This section presents the review design, search strategy, eligibility criteria, screening process, snowballing procedure, quality assessment, and data extraction approach used to construct the final evidence base.

3.1. Review Design

The study followed a PRISMA-informed systematic review process. PRISMA was used as a transparency framework for identification, deduplication, title/abstract/keyword screening, full-text eligibility assessment, citation searching, and qualitative synthesis [50]. The PRISMA 2020 checklist can be seen in the Supplementary Materials. The review is described as PRISMA-informed rather than a strict meta-analysis because the target literature is heterogeneous and includes conceptual, technical, empirical, qualitative, and design-oriented contributions. No statistical meta-analysis was appropriate.

Additional conceptual, normative, and methodological references were used in the Introduction and Background sections to define key concepts, including computational creativity, privacy-by-design, security-by-design, trustworthy AI, and threat modeling. These references were not counted as part of the PRISMA-derived primary synthesis unless they satisfied the eligibility criteria and were included in the final corpus.

Although the review protocol was not preregistered, the review process was documented in structured spreadsheets that recorded search results, screening decisions, exclusion reasons, data-extraction fields, and quality-assessment scores. To preserve the evidential consistency of the synthesis, only peer-reviewed and sufficiently developed studies were included as primary evidence. Other relevant but non-eligible contributions, such as preprints, review-style papers, viewpoints, or papers not specific to robotics, were retained only as background material when useful for framing or future directions.

3.2. Search Strategy and Sources

The search covered ACM Digital Library, Google Scholar, IEEE Xplore, Scopus, SpringerLink, and Web of Science (The bibliographic sources were accessed through the following platforms: ACM Digital Library, Google Scholar, IEEE Xplore, Scopus, SpringerLink, and Web of Science. The search was restricted to publications from 2020 to 2026). The search was restricted to publications from 2020 to 2026, to focus on recent developments in cybersecurity, privacy, AI-enabled robotics, and human–robot interaction. These sources were selected to capture work across computer science, human–robot interaction, cybersecurity, privacy engineering, robotics, artificial intelligence, and creative technologies. The initial search included the query: Cybersecurity AND "Creative robotics". Additional query variants were used across databases to capture adjacent terminology, including creative robotics, Co-Creative Robotics, social robotics, human–robot interaction, embodied AI, privacy-preserving robotics, security-by-design, threat modeling, blockchain provenance, federated learning, and cyber-physical security.

The search was intentionally broad because the intersection of cybersecurity, privacy and CCR is not yet consolidated. Relevant evidence may appear under labels such as social robots, collaborative robots, embodied AI, human–robot collaboration, AI-assisted design, visual privacy, digital twins, or autonomous cyber-physical systems. Broad retrieval was balanced by conservative screening and full-text eligibility criteria.

3.3. Eligibility Criteria

As reported in Table 1, the inclusion and exclusion criteria were designed to avoid two opposite risks: excluding transferable evidence too early, or including technically related but conceptually irrelevant work. During screening, a paper was retained only if it contributed to robotics, HRI, embodied AI, creative robotics, co-creative systems, or a clearly transferable security/privacy method. During full-text eligibility assessment, the publication window, peer-review status, contribution depth, and relevance to the RQs were checked. The primary publication window was 2020–2026. Older foundational sources were used only as conceptual background and were not counted as primary studies.

Table 1. Inclusion and exclusion criteria used during screening, full-text eligibility assessment and snowballing.

Type	ID	Criterion
Inclusion	IC1	Peer-reviewed journal, conference or workshop publication.
Inclusion	IC2	Written in English.
Inclusion	IC3	Published between 2020 and 2026, inclusive. Conceptual background references used only for definitions were exempt from this criterion and were not counted as primary evidence.
Inclusion	IC4	Focus on robotics, HRI, social robots, collaborative robots, creative robotics, Co-Creative Robotics, embodied AI or human-facing robotic systems.
Inclusion	IC5	Explicitly discusses cybersecurity, privacy, data protection, secure design, privacy-preserving mechanisms, trust, consent, provenance, accountability, autonomy, ethics or safety-security interaction.
Inclusion	IC6	Provides original research: model, method, framework, architecture, empirical study, prototype, experiment, case study, dataset, evaluation or design guidelines.
Inclusion	IC7	Provides information useful for at least one sub-research question.
Exclusion	EC1	Outside the 2020–2026 publication window, except for background-only conceptual or normative sources.
Exclusion	EC2	Not related to robotics, HRI, embodied AI, creative robotics, co-creative systems or human-facing robotic systems.
Exclusion	EC3	Mentions cybersecurity or privacy only superficially.
Exclusion	EC4	Mentions robotics only superficially.
Exclusion	EC5	Focuses only on cybersecurity education or computing curricula without relevant robotic/HRI/security-for-robots content.
Exclusion	EC6	Concerns robotic process automation or software bots rather than embodied or human-facing robots.
Exclusion	EC7	Review, survey, editorial, preface, monograph, demo-only paper, poster abstract or insufficiently developed contribution.
Exclusion	EC8	Not peer-reviewed or full text not accessible.
Exclusion	EC9	Unclear objectives, insufficient methodological detail or no identifiable contribution relevant to the RQs.

Table 1. *Cont.*

Type	ID	Criterion
Exclusion	EC10	Duplicate already represented by another record.
Exclusion	EC11	Fewer than six pages, excluding references and appendices, and insufficient methodological, technical, empirical or conceptual detail.

3.4. Screening and Deduplication

The initial search identified 623 records across the six sources: ACM Digital Library (n = 20), Google Scholar (n = 16), IEEE Xplore (n = 18), Scopus (n = 276), SpringerLink (n = 285), and Web of Science (n = 8). Duplicate detection was based primarily on DOI matching; where DOI information was missing or inconsistent, titles were normalized and compared. Five duplicates were removed, leaving 618 records for title, abstract, keyword, and metadata screening. Source-specific counts before and after deduplication are reported in Table 2.

Table 2. Record counts by source after identification and deduplication.

Source	Initial Records	Duplicates Removed	Records After Deduplication
ACM Digital Library	20	0	20
Google Scholar	16	0	16
IEEE Xplore	18	4	14
Scopus	276	0	276
SpringerLink	285	1	284
Web of Science	8	0	8
Total	623	5	618

The preliminary screening process was organized to reduce individual selection bias. Records were divided among the authors, and each record was screened by at least two authors based on title, abstract, keywords, and available metadata. Records were excluded at this stage only when they clearly fell outside the scope defined by the eligibility criteria in Table 1, for example, because they were unrelated to robotics, HRI, embodied AI, creative robotics, cybersecurity, privacy, trust, consent, accountability, autonomy, or safety-security interaction. Cases considered ambiguous by one or both screeners were discussed with a third author. When uncertainty remained, the record was retained for full-text assessment rather than excluded during preliminary screening.

As shown in Figure 1, 596 records were excluded during title, abstract, keyword, and metadata screening. This left 22 records for full-text retrieval and eligibility assessment. This conservative screening strategy was adopted to avoid excluding potentially relevant studies too early, given that the intersection between Co-Creative Robotics, cybersecurity, and privacy is emerging and inconsistently labeled across databases.

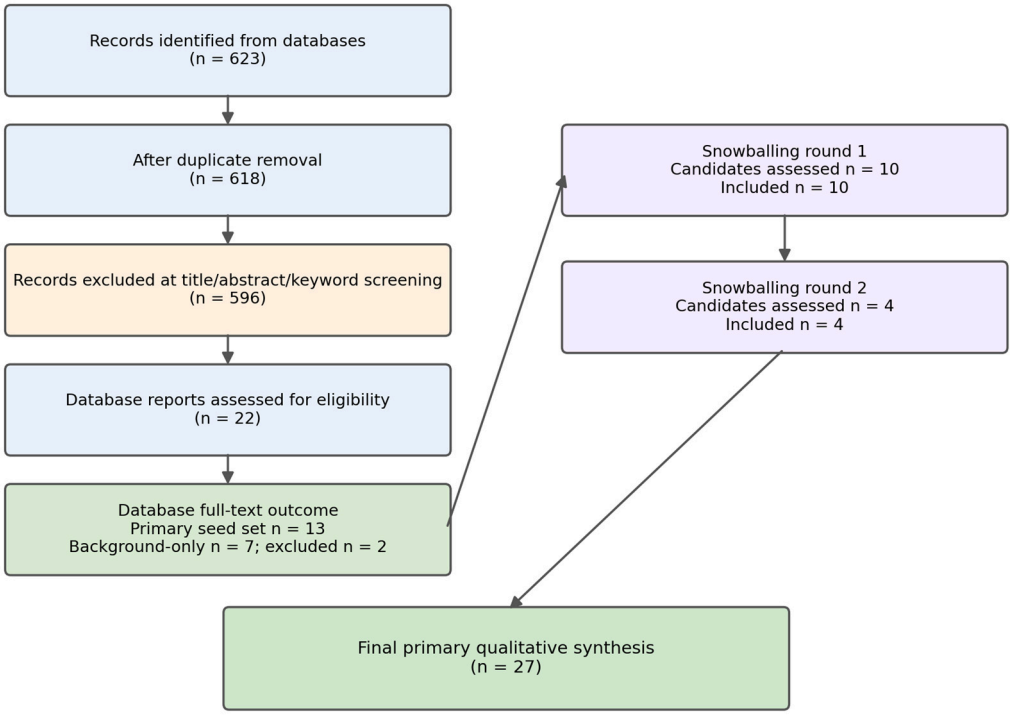


Figure 1. PRISMA-informed selection flow, including two focused backward/forward snowballing iterations.

3.5. Full-Text Eligibility Assessment and Snowballing

The 22 database-derived full texts retained after preliminary screening were assessed using the eligibility criteria reported in Table 1. As in the previous phase, the assessment was distributed among the authors so that each full text was independently reviewed by at least two authors. For each paper, the reviewers checked peer-review status, publication length, methodological or technical contribution, robotics/HRI relevance, and explicit connection to cybersecurity, privacy, trust, consent, accountability, autonomy, ethics, or safety-security interaction. Individual decisions were then compared and merged. Disagreements or uncertain cases were discussed with a third author until a shared decision was reached.

After full-text reading, papers were classified into three categories. Included papers satisfied all eligibility criteria and formed the database-derived primary corpus. Background-only papers were retained when they provided useful conceptual, normative, methodological, or technological framing for the review, but did not meet all criteria for primary synthesis. This occurred, for example, when a paper was review-style, a viewpoint, a preprint, non-robotics-specific, or only indirectly related to the research questions, yet still useful for contextualizing concepts such as security-by-design, privacy-preserving learning, provenance, multi-robot governance, or care-robot cybersecurity. Excluded papers differed: they were not used as either primary evidence or background framing because, after full-text assessment, they failed to meet minimum eligibility requirements, such as peer-reviewed publication status, sufficient depth of contribution, accessibility, or direct relevance to the review's scope. This distinction ensured that background-only papers supported conceptual framing without affecting the primary synthesis, PRISMA counts, quality assessment, or data extraction matrix.

To make the boundary between these analytical roles explicit, Table 3 indicates whether each full-text category contributed to the primary synthesis, quality assessment, and data extraction. As reported in Table 3, this process resulted in 13 database-derived studies included in the primary synthesis, 7 papers retained as background-only contributions,

and 2 papers excluded after full-text assessment. The 13 included studies were then used as the seed set for backward and forward snowballing, following established guidance on snowballing in the literature reviews [14].

Table 3. Updated study selection and full-text eligibility outcome.

Selection Step	n	Role in the Review
Database records identified	623	Initial search across six sources.
Duplicates removed	5	DOI and normalized-title matching.
Database records screened	618	Title, abstract, keywords and metadata.
Records excluded at screening	596	Not relevant or not eligible under IC/EC.
Database full texts assessed	22	Retained after Phase 2.
Database primary inclusions	13	Primary corpus; quality assessment, data extraction, thematic synthesis, and snowballing seed set.
Database background-only	7	Contextual framing only; not included in quality assessment, data extraction, or thematic synthesis.
Database full-text exclusions	2	Not eligible; used neither as primary evidence nor as background framing.
Snowballing round 1 candidates assessed	10	Backward and forward citation searching from the 13 seed studies.
Snowballing round 1 inclusions	10	All met full-text criteria.
Snowballing round 2 candidates assessed	4	Restricted citation searching from newly included studies.
Snowballing round 2 inclusions	4	All met full-text criteria.
Final primary synthesis	27	13 database inclusions + 14 snowballing inclusions.

Each seed paper was assigned to two authors, who independently inspected backward references and forward citations. The candidate lists produced by the two reviewers were then merged, deduplicated, and checked against the same eligibility criteria used for the database-derived records. Ambiguous snowballing candidates were discussed among the reviewers and, when needed, with a third author.

Snowballing was deliberately restrictive and included an initial pre-screening step before full-text eligibility assessment. Backward references and forward citations from the seed studies were first inspected based on title, abstract, venue, year, page count, and apparent topical relevance. Only records that clearly addressed robot or social-robot privacy, consent, disclosure, user control, privacy-compliant design, privacy-aware motion planning, or privacy-preserving multi-robot coordination were retained as full-text snowballing candidates. Records outside the 2020–2026 window, preprint-only records, review-style records, and non-robotic smart-home, IoT, legal, or general privacy papers were not retained for full-text assessment.

Thus, the ten candidates assessed in the first snowballing round and the four candidates assessed in the second round did not represent all backward references and forward citations identified from the seed studies. They represented only the records that had passed the restrictive pre-screening described above. The reported 100% inclusion rate, therefore, refers to this pre-filtered set of full-text snowballing candidates.

As summarized in Table 3, the final primary corpus comprised 27 studies: 13 identified through a database search and 14 added through two controlled snowballing iterations. Seven additional papers were retained as background-only contributions, and two database-

derived full texts were excluded after eligibility assessment. Table 3 also distinguishes the analytical role of each full-text category in the review.

3.6. Quality Assessment and Data Extraction

Quality assessment was applied to all 27 primary studies included in the final synthesis. Because the corpus was methodologically heterogeneous, including empirical HRI studies, technical cybersecurity papers, formal or modeling contributions, design-oriented studies, and conceptual frameworks, the assessment was designed to evaluate transparency, relevance, and evidential usefulness rather than impose a single experimental standard.

Each study was assessed using seven criteria: clarity of research aims (QA1), clarity of the robotic, HRI, embodied, or co-creative context (QA2), explicit treatment of cybersecurity, privacy, trust, ethics, autonomy, or agency (QA3), level of methodological or technical detail (QA4), presence of empirical, technical, conceptual, formal, or design-based evidence (QA5), discussion of creativity, autonomy, agency, user control, or interaction quality (QA6), and discussion of limitations, risks, or open challenges (QA7). Each criterion was scored as 1 when fully satisfied, 0.5 when partially satisfied, and 0 when not satisfied. The maximum score was 7. Quality levels were defined as High ($\geq$6), Medium (4.5–5.9), and Low (<4.5).

The quality assessment was conducted independently by at least two authors for each included study, using the seven criteria reported above. Initial scores were compared, and disagreements or borderline cases were discussed until consensus was reached. When needed, a third author was involved in resolving interpretive differences by checking the assessment against the extracted data and eligibility criteria. We did not compute a formal inter-rater agreement coefficient because the quality assessment was used as a consensus-based evidence-weighting step rather than as an exclusion mechanism. The quality score was not used as an automatic exclusion criterion after eligibility was established. Instead, it was used to weigh the contribution of each paper in the thematic synthesis: high-quality studies were treated as stronger evidence, whereas medium-quality studies were used more cautiously, especially when they were conceptual, exploratory, or only indirectly transferable to CCR.

Data extraction was performed using a structured matrix. For each study, the extracted fields included bibliographic information, study type, application domain, robotic or AI context, cybersecurity focus, privacy focus, ethical or governance focus, evidence of creativity/autonomy, lifecycle phase, covered research questions, methodological approach, main findings, limitations, and role in the synthesis. The extraction was intended not only to identify security and privacy mechanisms, but also to determine whether and how those mechanisms could be interpreted as compatible with the creative and autonomous properties of CCR.

The extracted data were then used to organize the thematic synthesis by research question. Studies that directly address CCR or creative-HRI were used to define the creative properties to be preserved. Social-robot privacy studies were used to derive requirements for consent, disclosure, user control, privacy settings, privacy-aware behavior, and embodied privacy. Cybersecurity, security modeling, digital twin, and threat modeling studies were used as transferable evidence for lifecycle protection and assurance. Ethical and governance-oriented studies were used to interpret accountability, contextual autonomy, inclusion, and responsible deployment.

4. Results

This section first describes the final corpus and its quality profile, and then synthesizes the evidence by research question to identify how cybersecurity-by-design and privacy-by-design can be integrated into CCR while preserving creative autonomy and user agency.

4.1. Composition of the Final Corpus

Table 4 provides a study-level summary of all 27 primary studies included in the qualitative synthesis, ensuring consistency with the final number of studies reported in the PRISMA flowchart. The corpus does not yet form a mature, self-contained literature on cybersecurity and privacy in CCR. Instead, it represents four complementary groups of evidence with different levels of proximity to CCR.

Table 4. Summary of the 27 primary studies included in the qualitative synthesis.

ID	Study (Reference)	Study Design/Method	Domain or Context	Main Relevance to the Review
P01	Nikghadam-Hojjati et al. (2024) [1]	Conceptual study/focused brainstorming	Co-Creative Robotics	Defines CCR and frames creativity, autonomy, ethics, and cybersecurity as core concerns.
P02	Ali et al. (2021) [2]	Randomized empirical HRI study	Social robots and creativity	Shows that embodied social robots can scaffold and elicit verbal, figural, and constructional creativity.
P03	Song et al. (2022) [3]	Empirical design experiment	AI-assisted engineering design	Shows how AI assistance affects design exploration, novelty, effectiveness, and human initiative under complexity.
P04	Callari et al. (2024) [10]	Delphi-based ethical framework	Human–robot collaboration in manufacturing	Supports accountability, dignity, worker agency, and human-centered governance in collaborative robotics.
P05	Kayan et al. (2024) [6]	Technical empirical study	Industrial robotic arms	Provides cyber-physical anomaly detection evidence beyond network-only monitoring.
P06	Roberts et al. (2021) [7]	Cyber range/testbed evaluation	Autonomous vehicle shuttles	Demonstrates safe, repeatable testing of attacks and mitigations for autonomous robotic systems.
P07	Zhang et al. (2024) [11]	User-centered design and user study	Accessible visual privacy	Provides user-controlled obfuscation principles transferable to robot cameras and multimodal sensing.
P08	Ghoshal et al. (2025) [12]	Conceptual ethics analysis	Embodied AI and marginalized contexts	Highlights contextual autonomy, spatial justice, inclusion, and limitations of universal governance assumptions.
P09	Spencer et al. (2024) [42]	Security modeling framework	Socio-technical security modeling	Supports participatory modeling of assets, harms, stakeholders, referent objects, and security stance.
P10	Arora and Hastings (2025) [43]	Multilayer security framework	Agentic AI systems	Provides lifecycle security concepts for accountable autonomous and agentic systems.

Table 4. *Cont.*

ID	Study (Reference)	Study Design/Method	Domain or Context	Main Relevance to the Review
P11	Shaikh et al. (2023) [8]	Formal verification/model checking case study	Digital twin security	Supports architecture-level analysis of attack success and countermeasure cost in cyber-physical systems.
P12	Thompson et al. (2024) [9]	Interview study	Medical-device threat modeling	Provides evidence on secure-by-design practices and safety-security trade-offs in regulated cyber-physical systems.
P13	Trentesaux and Karnouskos (2020) [13]	Conceptual and case-based analysis	Autonomous intelligent CPS	Supports lifecycle responsibility, machine ethics, and accountability for autonomous cyber-physical behavior.
P14	Shome et al. (2023) [15]	Technical motion-planning study	Privacy in robot motion planning	Shows that robot trajectories and sensing configurations can preserve, ignore, or violate privacy.
P15	Lutz and Tamò-Larrieux (2020) [16]	Survey/structural equation modeling	Social robots	Explains privacy concerns, perceived benefits, trust, and intention to use social robots.
P16	Dietrich et al. (2023) [17]	Vignette-based empirical study	Robot-mediated disclosure	Shows that privacy-appropriate disclosure depends on context, relationship, purpose, and information type.
P17	Tang et al. (2022) [18]	Privacy controller and user studies	Privacy-aware social robots	Provides rules and conversational metadata mechanisms for privacy-aware social-robot disclosure.
P18	Lutz and Tamò-Larrieux (2021) [19]	Experimental vignette study	Social robot acceptance	Shows that privacy-friendly robot design can influence intention to use social robots.
P19	Aryania et al. (2025) [20]	Experimental user study	Public-space social robots	Examines how privacy warnings affect acceptance and actual disclosure of personal information.
P20	Yang et al. (2022) [21]	User study	Privacy-sensitive robot behaviors	Shows how gaze, distance, and clarity of intent can communicate privacy-respecting behavior.
P21	Stapels et al. (2023) [22]	Two online experiments	Robot privacy settings	Shows how strict settings and user choice affect trust, ambivalence, likeability, and self-disclosure.
P22	Windl et al. (2024) [23]	Survey, focus groups, and large-scale survey	Domestic robots	Provides communication patterns for camera, microphone, and connectivity states in domestic robots.
P23	Alsayegh et al. (2022) [24]	Technical method with physical proof-of-concept	Multi-robot task allocation	Demonstrates secure multi-party computation for private coordination among robots.

Table 4. *Cont.*

ID	Study (Reference)	Study Design/Method	Domain or Context	Main Relevance to the Review
P24	Jayaraman et al. (2024) [25]	Empirical study	Social robots in healthcare	Shows how privacy and utility perceptions vary with task and healthcare context.
P25	Horstmann et al. (2020) [26]	Use cases, GDPR analysis, and app mock-up	Private-household social robots	Translates GDPR-oriented privacy requirements into concrete technical and interaction controls.
P26	Sakai et al. (2023) [27]	HAI experiments	Robot notification and permission	Shows effects and limits of robot notifications for permission and personal-information use.
P27	Levinson et al. (2024) [28]	Qualitative co-learning study	Family social robots in the home	Highlights multi-stakeholder privacy, child/parent agency, and family boundary management.

- The first group contains core CCR or creative-HRI studies that define what must be preserved: creativity, creative scaffolding, human initiative, co-creative exploration and design autonomy [1–3]. These studies provide the most direct evidence for the creative and interactional properties of CCR.
- The second group contains related robotics, HRI and embodied-AI studies addressing human-centered manufacturing, industrial robotic security, cyber-physical testbeds, embodied governance and autonomous CPS ethics [6,7,10,12,13]. These studies provide adjacent evidence on embodied autonomy, human-centered interaction, and cyber-physical risk.
- The third group contains transferable security and assurance methods, including security modeling, agentic-AI security, digital-twin verification and medical-device threat modeling [8,9,42,43]. These studies are not treated as direct CCR evidence, but as sources of security-by-design and assurance mechanisms that require adaptation to co-creative contexts.
- The fourth group contains direct evidence on social-robot privacy, including privacy-aware motion planning, privacy paradoxes, disclosure, privacy controllers, warnings, settings, sensor-state communication, healthcare contexts, domestic privacy controls, permission mechanisms and family boundary management [15–28]. These studies provide interaction-level evidence on privacy, consent, disclosure, trust, and user agency in embodied or human-facing robotic systems.

Accordingly, the synthesis does not assume that findings from adjacent or transferable domains apply to CCR without modification. Rather, these studies are used to identify mechanisms, risks, and design principles that must be interpreted in light of CCR's specific requirements for creative autonomy, open-ended interaction, and human agency.

4.2. Quality Assessment

The quality profile of the final corpus supports its use for qualitative and thematic synthesis. As reported in Table 5, twenty-one of the 27 primary studies were rated High, six were rated Medium, and none were rated Low. The average score was 6.19 out of 7. This distribution indicates that the corpus is sufficiently documented in terms of research aims, methodological or technical detail, evidence type, relevance to robotics/HRI or embodied systems, and discussion of limitations or open challenges.

Table 5. Quality assessment summary for the 27 primary studies.

Quality Level	n	Interpretation
High	21	Strong empirical, technical, formal or framework contribution; clear aims and usable evidence.
Medium	6	Relevant but more conceptual, indirect, exploratory or transfer-oriented; useful with cautious interpretation.
Low	0	No primary study fell below the quality threshold.
Average score	6.19/7	Quality sufficient for narrative and thematic synthesis.

No study was excluded based on quality score after meeting the eligibility criteria. Instead, the quality assessment was used to calibrate the evidential weight assigned to each contribution. High-scoring studies typically provided empirical HRI evidence, technical validation, formal analysis, design evaluation, or well-developed framework contributions. Medium-scoring studies were generally more conceptual, exploratory, or transfer-oriented; they remained useful for synthesis, but their findings were interpreted with greater caution.

The main limitation of the quality profile is not evidential weakness, but directness. Only a subset of studies explicitly addresses CCR or creative-HRI. Many studies focus instead on social robots, domestic robots, healthcare robots, industrial robots, autonomous vehicles, multi-robot systems, privacy engineering, or security-engineering contexts. These studies were interpreted as transferable evidence only when the underlying risk, method, or design mechanism was applicable to co-creative embodied interaction. This approach prevents overclaiming while still allowing the review to derive lifecycle-oriented implications for secure and privacy-aware CCR.

4.3. SRQ1—Challenges, Issues and Opportunities

As summarized in Table 6, the evidence for SRQ1 can be organized into five partially overlapping clusters of challenges and opportunities. These clusters are not mutually exclusive because several studies address more than one issue, such as privacy, trust, consent, autonomy, embodied sensing, user control, or cyber-physical risk. Therefore, the counts reported in the table indicate thematic coverage within the final corpus rather than mutually exclusive categories.

Table 6. Distribution of primary studies across SRQ1 challenge and opportunity clusters.

SRQ1 Cluster	Main Focus	Primary Studies Contributing to the Cluster	n
Multimodal privacy exposure	Sensing, cameras, microphones, connectivity states, visual privacy, privacy-sensitive situations, privacy-aware motion, and contextual data capture	[11,15,21–23,25,26]	7
Disclosure and boundary management	Information sharing, conversational privacy, consent, family privacy, contextual disclosure, and co-ownership of information	[17,18,20,26–28]	6
Privacy-utility-trust trade-off	Adoption, perceived benefit, privacy paradox, trust, use intention, privacy settings, warnings, agreement, and healthcare utility	[16,18–20,22,25,27]	7

Table 6. *Cont.*

SRQ1 Cluster	Main Focus	Primary Studies Contributing to the Cluster	n
Cyber-physical manipulation and embodied risk	Cyber-physical attacks, anomaly detection, test ranges, digital twins, privacy-aware motion planning, and privacy-preserving multi-robot coordination	[6–8,15,24]	5
Autonomy, accountability, and inclusion	Creative autonomy, user agency, human-centric governance, postcolonial ethics, agentic AI, autonomous CPS, family agency, and responsible interaction	[1–3,10,12,13,28,42,43]	9

The distribution shows that SRQ1 is supported by both technical and interactional evidence. The most represented cluster concerns autonomy, accountability, and inclusion, followed by multimodal privacy exposure and privacy-utility-trust trade-offs. This pattern is consistent with the nature of CCR as an embodied and interactional field: cybersecurity and privacy risks are not limited to data leakage or external attacks, but also concern how users understand, negotiate, and contest the robot's sensing, memory, disclosure, movement, action, and influence during collaborative activity.

- The first cluster is multimodal privacy exposure. CCR systems may process images, audio, gestures, gaze, proximity, creative artifacts, design traces, and behavioral cues. The privacy problem is not limited to data leakage; it also concerns whether users understand what the robot senses, how data is transformed, and who controls obfuscation, retention, or reuse. Accessible visual obfuscation shows that agency-preserving privacy support requires users to remain involved in obfuscation decisions rather than being replaced by opaque automation [11]. Social-robot privacy studies add that robots should communicate camera, microphone, and connectivity states [23], exhibit privacy-respecting behaviors such as turning away or moving away in sensitive situations [21], and make privacy settings understandable and controllable [22].
- The second cluster is disclosure and boundary management. Social robots may mediate information between people, overhear conversations, remember personal information, or repeat it in a different context. CONFIDANT shows that conversational metadata can be used to infer privacy boundaries and improve perceived privacy awareness, trustworthiness, and social awareness [18]. Work on privacy-appropriate disclosure shows that disclosure cannot be reduced to a generic share/do-not-share decision; appropriateness depends on relationship, context, purpose, and information type [17]. Privacy warnings and robot notifications further show that users' understanding of data use and permission decisions depends on how robots communicate data practices during or after interaction [20,27]. Family-oriented privacy research adds that privacy can be negotiated among multiple stakeholders, such as parents and children, who may differ in agency, responsibility, and ownership of information [28].
- The third cluster is the privacy-utility-trust trade-off. The robot privacy paradox shows that perceived benefits can override privacy concerns in social-robot adoption decisions [16]. A later experimental vignette study indicates that privacy-friendly robot design can increase intention to use [19]. Studies on privacy settings show that strict settings and user choice can influence trust, ambivalence, likeability, and willingness to self-disclose [22]. Healthcare evidence adds that the task and context in which a robot operates may matter more than the robot's human-like appearance

when users evaluate privacy and utility [25]. Together, these studies suggest that privacy controls in CCR should be evaluated not only as compliance mechanisms but also for their effects on perceived usefulness, trust, and willingness to collaborate.

- The fourth cluster is cyber-physical manipulation and embodied risk. Robotic embodiment transforms cyberattacks into physical consequences. CASPER demonstrates that network-based intrusion detection is insufficient when attacks manipulate an actuator's behavior or physical motion; independent motion-based anomaly detection can provide a secondary protection layer for industrial robotic arms [6]. Cyber ranges for autonomous shuttles similarly show that safe and repeatable cybersecurity testing is needed before exposing real systems to edge and corner cases [7]. Privacy-aware motion-planning work extends this point to privacy: even legitimate motion and sensing can be configured to preserve, ignore, or violate privacy [15]. Privacy-preserving multi-robot task allocation further shows that coordination mechanisms can be designed to protect private information while still enabling distributed robotic action [24].
- The fifth cluster concerns autonomy, accountability, and inclusion. Co-creative robots may make suggestions, scaffold ideas, adapt their behavior, and influence human choices. This is valuable for creativity, but it complicates responsibility. HRC ethics research highlights accountability, human-centric governance, and the need to preserve the worker or user as an active participant rather than an object of automation [10]. Postcolonial analysis of embodied AI shows that governance frameworks may impose assumptions about autonomy, testing infrastructure, and individual decision-making that do not transfer across sociotechnical contexts [12]. Security modeling and agentic AI work further indicate that autonomy requires explicit reasoning about assets, stakeholders, control, accountability, and lifecycle risk [42,43]. In family-oriented social-robot privacy, autonomy and control are also distributed across stakeholders, showing that agency may be negotiated rather than individually held [28].

The opportunity is that cybersecurity and privacy can become trust-enabling design resources. If users understand the boundaries of sensing, if control remains adjustable and contestable, if disclosure rules are context-sensitive, if motion is privacy-aware, if creative outputs are traceable, and if robots can be tested and monitored without suppressing interaction, security and privacy can increase confidence and willingness to engage creatively.

Answer to SRQ1. The strongest evidence indicates that CCR cannot treat cybersecurity and privacy as external constraints added after creative interaction has been designed. The main recurring risks concern multimodal and ambient data capture, disclosure and boundary management, cyber-physical manipulation, and gaps in agency and accountability. The main uncertainty is how these risks scale in genuinely open-ended CCR settings, where creative deviation, improvisation, and ambiguity are not exceptions but expected forms of interaction. A key research gap is therefore the development of protection mechanisms that build trust and preserve agency without narrowing the user's creative exploration space.

4.4. SRQ2—Application Domains

As summarized in Table 7, the evidence for SRQ2 covers several application domains and transferable contexts. The domain groups are partially overlapping because some studies address both a concrete robotic setting and a transferable security or privacy mechanism. Accordingly, the counts indicate domain coverage within the final corpus rather than mutually exclusive classes.

Table 7. Distribution of primary studies across SRQ2 application domains.

Application Domain Group	Main Focus	Primary Studies Contributing to the Domain Group	n
Core CCR, creative-HRI, and AI-assisted design	Co-Creative Robotics, social robots for creativity, creative scaffolding, and AI-assisted engineering design	[1–3]	3
Social, domestic, and family robots	Household robots, robot-mediated disclosure, privacy settings, privacy-sensitive behaviors, sensor-state communication, privacy-compliant domestic robots, and family privacy	[15–19,21–23,26,28]	10
Healthcare, rehabilitation, and public interaction	Healthcare robots, home rehabilitation, public-space interaction, privacy warnings, permission, and disclosure of personal information	[20,25,27]	3
Manufacturing, industrial robotics, and autonomous mobility	People-centric manufacturing, industrial robotic arms, cyber-physical anomaly detection, autonomous shuttles, and cybersecurity validation	[6,7,10]	3
Multi-robot coordination	Privacy-preserving multi-robot task allocation and coordination without exposing private valuations or assignments	[24]	1
Transferable security, privacy, and assurance domains	Accessible visual privacy, digital twins, medical-device threat modeling, security modeling, and agentic-AI security	[8,9,11,42,43]	5
Embodied AI, autonomous CPS, and governance	Embodied AI ethics, postcolonial robotics, autonomous intelligent cyber-physical systems, contextual autonomy, and lifecycle responsibility	[12,13]	2

The table shows that CCR cybersecurity and privacy cannot be associated with a single application domain. Core CCR and creative-HRI studies define the creative and collaborative settings in which protection mechanisms must preserve exploration, scaffolding, novelty, initiative, and meaningful human participation [1–3]. These studies are essential for defining the domain, but they do not yet constitute a mature application literature on cybersecurity and privacy for CCR.

A substantial part of the evidence comes from social, domestic, and family robots. These studies address household robots, privacy-sensitive situations, domestic sensor-state communication, privacy settings, robot-mediated disclosure, family discussions about robots in the home, and privacy-compliant social robots for private households [15–19,21–23,26,28]. Although they are not always labeled as creative robotics, their findings are directly relevant to CCR because co-creative systems often operate in intimate, expressive, and multi-stakeholder settings where sensing, memory, disclosure, and user control influence trust and agency.

Healthcare, rehabilitation, and public interaction form another important domain group. Healthcare evidence shows that perceptions of privacy and utility depend strongly on the setting and task, especially in home healthcare and rehabilitation [25]. Public-space and permission-oriented studies show that privacy warnings and robot notifications can

affect users' understanding, consent, and disclosure behavior [20,27]. These domains matter for CCR because co-creative robots may be deployed in education, museums, therapy, rehabilitation, public engagement, and care contexts where creativity, data sensitivity, and social vulnerability coexist.

Related cyber-physical robotic domains include people-centric manufacturing, industrial robotic arms, autonomous vehicles, and multi-robot coordination. In manufacturing, human–robot collaboration raises ethical and organizational issues around monitoring, dignity, role allocation, accountability, and future work [10]. In industrial robotic arms, the focus is on cyber-physical integrity and anomaly detection [6]. In autonomous shuttles, the domain is safety-critical mobility and cybersecurity validation [7]. In multi-robot systems, secure multi-party computation demonstrates how coordination can occur without exposing private valuations, utilities, or assignments [24].

Transferable domains include accessible visual privacy, digital twins, medical device threat modeling, security modeling, agentic AI, embodied AI ethics, and autonomous cyber-physical systems [8,9,11–13,42,43]. These are not CCR domains in themselves, but they address lifecycle problems CCR will face: sensing privacy, participatory security, autonomous agent control, virtual assurance, safety-security integration, contextual autonomy, and lifecycle responsibility.

Answer to SRQ2. The evidence is strongest in adjacent domains rather than in mature CCR-specific deployments. Social robots, healthcare robots, domestic robots, public-space robots, people-centric manufacturing, autonomous mobility, digital twins, and multi-robot coordination provide useful analogs for CCR, especially because they involve embodied sensing, user trust, and situated interaction. However, the main gap is the limited number of studies that evaluate cybersecurity and privacy directly in physical co-creative robots. Future work should therefore move from domain transfer toward CCR-specific validation in creative, educational, artistic, workplace, and public settings.

4.5. SRQ3—Methods and Technical Approaches

Table 8 summarizes the main method families identified in the final corpus. The methods are not mutually exclusive: several studies combine interaction design, privacy protection, user control, security modeling, or cyber-physical assurance. The table shows the range of methodological contributions that can support cybersecurity-by-design and privacy-by-design in CCR.

Table 8. Method families emerging from the synthesis.

Method Family	Description	Relevance to CCR
Creativity scaffolding	Robot demonstrates ideas, asks questions, or provides creative prompts.	Preserves creative initiative while enabling robotic contribution [2].
AI-assisted design support	AI recommends or explores design alternatives.	Supports exploration while monitoring overreliance and loss of initiative [3].
Privacy controllers	Conversational metadata and rules classify information sensitivity.	Controls disclosure boundaries in social and co-creative conversations [18].
Privacy-appropriate disclosure	Robot mediates what information is shared between people.	Supports context-sensitive disclosure and social trust [17].
Privacy settings and user choice	Strict or user-selectable settings change trust and self-disclosure.	Makes protection adjustable rather than hidden [22].

Table 8. *Cont.*

Method Family	Description	Relevance to CCR
Privacy-aware robot behaviors	Gaze, distance, and clarity of intent make privacy protection visible.	Translates privacy into embodied interaction patterns [21].
Warnings and notifications	Robot reminders and warnings inform users about data use.	Improve understanding but do not automatically reduce disclosure [20,27].
Communication patterns	Robot communicates microphone, camera, and connectivity states.	Makes sensing status interpretable in domestic environments [23].
Privacy-aware motion planning	Cost functions can preserve, ignore, or violate privacy.	Protects privacy at the level of physical motion and sensing [15].
Secure multi-party computation	Robots allocate tasks without revealing private valuations or assignments.	Enables privacy-preserving distributed and multi-robot coordination [24].
Privacy-preserving obfuscation	Users control visual privacy transformations.	Transferable to robot cameras and multimodal sensing [11].
Cyber-physical anomaly detection	Independent sensing of motion anomalies.	Protects embodied behavior beyond network-layer security [6].
Cyber ranges and digital twins	Controlled attack testing and formal model checking.	Validate cyber-physical and architecture-level protections [7,8].
Security and threat modeling	Structured elicitation of assets, harms, trust boundaries, and mitigations.	Supports secure-by-design and participatory CCR development [9,40,42,43].

The distribution of methods in Table 8 shows that the evidence base does not point to a single dominant solution. Instead, it suggests a layered method stack.

- The first layer concerns creative support: creativity scaffolding and AI-assisted design clarify how robotic or AI contributions can stimulate exploration without replacing human initiative [2,3]. These methods define the creative function that protection mechanisms must preserve.
- The second layer concerns interaction-level privacy. Privacy controllers, privacy-appropriate disclosure, privacy settings, privacy-aware behaviors, warnings, notifications, and communication patterns show how privacy can be embedded directly into human–robot interaction [17,18,20–23,27]. This layer is especially relevant to CCR because co-creative robots may collect, remember, transform, and disclose creative traces during interaction. The evidence indicates that privacy protection should be visible, adjustable, and context-sensitive, rather than hidden only in backend data management procedures.
- The third layer concerns embodied and distributed robotic protection. Privacy-aware motion planning demonstrates that privacy can be affected by where and how a robot moves, not only by what data it stores [15]. Secure multi-party computation shows that multi-robot coordination can be performed without exposing private valuations or assignments [24]. Cyber-physical anomaly detection adds a complementary security mechanism by monitoring physical deviations that may not be visible at the network layer [6]. These methods are important for CCR because creative robots are embodied systems whose movements, sensing angles, proximity, and coordination choices can affect both privacy and safety.

- The fourth layer concerns assurance and lifecycle governance. Cyber ranges, digital twins, security modeling, and threat modeling provide methods for identifying risks, testing attacks and countermeasures, analyzing system architectures, and clarifying assets, harms, trust boundaries, and responsibilities [7–9,40,42,43]. These methods are not specific to CCR but are transferable, as CCR systems also require design-time risk analysis, controlled validation, runtime monitoring, and post-deployment accountability.

Across these layers, transferable methods should be interpreted through translational scenarios rather than applied mechanically. For creative-support methods, this means that AI assistance or robotic scaffolding should not narrow the user's exploration space or push the user toward overly safe, predictable, or system-preferred outputs. For interaction-level privacy methods, warnings, notifications, settings, and disclosure controls should be designed to inform users without repeatedly interrupting improvisation or fragmenting the creative flow. For embodied and distributed robotic protection, a motion-anomaly detector such as CASPER, originally validated on industrial robotic arms, would require adaptive thresholds in a painting, dance, or music-improvising robot, where an irregular trajectory, sub-optimal rhythm, or exaggerated gesture may be an intentional expressive act rather than evidence of compromise. Similarly, privacy-aware motion planning can be translated to a museum or gallery robot by selecting trajectories and sensor orientations that avoid unnecessary capture of bystanders while preserving the primary user's collaborative experience, while secure multi-party computation can support multi-robot creative installations without revealing participants' private preferences or creative choices. Finally, assurance methods such as cyber ranges, digital twins, and threat modeling should include scenarios that test not only attack resistance, but also whether protection mechanisms preserve ambiguity, exploration, embodiment, and human initiative.

Taken together, the methods reported in Table 8 indicate that CCR requires compositional integration rather than isolated controls. A privacy warning alone cannot address cyber-physical manipulation; anomaly detection alone cannot address disclosure boundaries; and security modeling alone cannot preserve creative agency unless it is connected to interaction design. Therefore, a credible cybersecurity-by-design and privacy-by-design approach for CCR should combine participatory threat modeling, privacy-preserving sensing, disclosure control, adjustable privacy settings, privacy-aware robot behavior, secure model and data handling, cyber-physical monitoring, safe testing, explainability, provenance, and post-deployment governance.

Answer to SRQ3. The strongest methodological contribution of the corpus is not a single control, but a set of transferable mechanisms that can be combined across the CCR lifecycle: privacy controllers, disclosure mechanisms, adjustable privacy settings, sensor-state communication, privacy-aware motion planning, visual obfuscation, anomaly detection, cyber ranges, digital twins, and threat modeling. The remaining uncertainty concerns how these mechanisms should be calibrated in creative contexts, where unusual movements, incomplete artifacts, or unexpected user inputs may be legitimate forms of exploration. The key research gap is the lack of CCR-specific translational scenarios and benchmarks for testing whether protection mechanisms preserve creativity, agency, and interaction quality.

4.6. SRQ4—Technical, Ethical and Regulatory Challenges

SRQ4 concerns the barriers that may affect the integration of cybersecurity and privacy into CCR. As summarized in Table 9, these barriers are not only technical. They involve cyber-physical robustness, interaction-level privacy, consent and disclosure, user agency,

multi-stakeholder governance, and the translation of regulatory principles into concrete design evidence.

Table 9. Technical, ethical and regulatory challenges emerging from the synthesis.

Challenge Area	Evidence from the Corpus	Implication for CCR
Cyber-physical integrity	Robotic attacks may affect motion, actuation, sensor interpretation, and physical safety [6,7].	CCR systems require protection at both cyber and physical layers.
Runtime validation and assurance	Digital twins, cyber ranges, and formal verification support attack testing and countermeasure assessment [7,8].	Creative autonomy should be tested under controlled conditions before deployment.
Privacy-aware sensing and motion	Privacy can be affected by where robots move, what they sense, and how sensor states are communicated [15,23].	Privacy must be designed into embodied behavior, not only into data storage.
Consent, notification, and disclosure	Warnings, robot notifications, privacy controllers, and disclosure strategies affect understanding, trust, and permission decisions [17,18,20,27].	CCR needs contextual and interactional consent mechanisms rather than one-time notices only.
User control and privacy settings	Privacy settings and user choice influence trust, self-disclosure, and perceived control [22].	Protection should remain adjustable and contestable by users.
Multi-stakeholder privacy	Family and domestic studies show that privacy may be negotiated among people with different roles, agency, and information ownership [26,28].	CCR should account for bystanders, collaborators, caregivers, children, and other affected stakeholders.
Accountability and human agency	HRC and embodied AI ethics highlight dignity, role allocation, inclusion, and accountability [10,12,13].	Security and privacy controls should not turn users into passive objects of automation.
Regulatory translation	GDPR-oriented design requires concrete mechanisms such as consent, deletion, access restriction, and data minimization [4,26].	Compliance must be converted into auditable design choices and lifecycle evidence.

The table shows that the main challenge is not the absence of individual mechanisms, but their integration across technical, interactional, ethical, and regulatory levels. A CCR system may include strong cybersecurity controls, but still undermine privacy if users cannot understand what the robot senses or discloses. Conversely, a system may provide privacy notices or settings but remain unsafe if cyber-physical attacks can alter motion, actuation, or sensor interpretation.

- Technical challenges include data integrity, sensor spoofing, adversarial manipulation, unsafe actuation, secure communications, runtime monitoring, model lifecycle risk, and validation of autonomous behavior. CASPER and the autonomous shuttle cyber range show that embodied robotic systems require protection at the physical and cyber layers simultaneously [6,7]. Digital-twin security analysis adds a formal route for

estimating attack success and countermeasure costs across layered architectures [8]. Privacy-aware motion planning extends this technical view by showing that privacy can be affected by where the robot moves and how it points its sensors, not only by downstream data storage [15].

- Privacy challenges include consent, sensing visibility, bystander privacy, disclosure to third parties, contextual appropriateness, user control, and the privacy-utility trade-off. The evidence warns against simplistic solutions. Privacy warnings can increase acceptance but may not reduce actual disclosure behavior [20]. Robot notifications can improve understanding and permission rates, but do not necessarily increase perceived safety or agreement, and adding a rationale may not produce the expected effect [27]. Privacy settings can influence trust, ambivalence, likeability, and willingness to self-disclose, but they also expose trade-offs between privacy, utility, and user burden [22]. Ambient privacy is particularly important for CCR because co-creative robots may operate in public or semi-public settings such as classrooms, museums, galleries, laboratories, workshops, or care environments. In these contexts, the robot may capture speech, faces, gestures, movements, proximity, affective cues, or creative traces from bystanders and ambient participants who have not entered the primary consent loop. This makes one-to-one consent insufficient. CCR systems should therefore distinguish between primary users, co-present collaborators, bystanders, and affected stakeholders, and should support mechanisms such as visible sensor-state communication, spatial privacy zones, edge-based background blurring, selective audio capture, local processing where possible, and policies for excluding or minimizing data from non-participating persons. This direction is technically plausible because privacy-preserving visual sensing and edge-based video analytics can already support task-oriented capture, on-device filtering, obfuscation, and transmission of only task-relevant visual information [51]. These mechanisms are necessary not only for compliance, but also to preserve trust in shared creative environments.
- Ethical challenges include agency, accountability, surveillance, exclusion, overtrust, over-constraint, and multi-stakeholder negotiation. In family settings, parents and children may disagree about what the robot should know or share, and the appropriate boundary may depend on who owns the information and who has responsibility in the situation [28]. In healthcare, users may perceive high utility and high privacy sensitivity simultaneously, especially in home-based care and rehabilitation [25]. In workplaces and manufacturing settings, privacy and security controls must also respect dignity, role allocation, accountability, and the active participation of workers or users [10].
- Regulatory challenges concern translation rather than principle alone. GDPR establishes data protection by design and by default [4], but CCR systems need concrete engineering choices: what data are collected, where they are processed, how long they are retained, how consent is represented during creative collaboration, and how users can inspect, contest, or redirect robot behavior. The private-household app study illustrates how GDPR-oriented requirements can be translated into design controls such as explicit consent, microphone control, deletion of conversation data, room access restrictions, sleep mode, and adjustable robot curiosity [26]. Security-by-design similarly requires more than compliance language; it requires threat modeling, architecture-level reasoning, testing, monitoring, and documented accountability [5,9,42,43].

The central cross-cutting challenge is proportionality. Controls must be strong enough to prevent harm, but flexible enough to preserve creative exploration. Excessive monitoring may protect the system while undermining privacy. Excessive restrictions may prevent misuse but also suppress improvisation. Insufficient controls may preserve apparent freedom

while eroding trust and exposing users to hidden risks. For CCR, the design problem is not to choose between protection and autonomy, but to make protection adjustable, explainable, testable, and compatible with human creative agency.

Answer to SRQ4. The strongest finding is that transparency and user control are necessary but not sufficient conditions for trustworthy CCR. The evidence on warnings, notifications, privacy settings, and disclosure controls is mixed: these mechanisms can improve awareness or acceptance, but they do not automatically reduce disclosure, increase perceived safety, or remove user burden. The main unresolved challenge is therefore to design privacy and security mechanisms that are not only understandable but also behaviorally effective, proportionate, and usable in embodied co-creative interaction. This is especially important for ambient and bystander data capture, where consent, sensing visibility, and accountability extend beyond the primary user.

4.7. SRQ5—Lifecycle-Oriented Future Directions

SRQ5 concerns how cybersecurity-by-design and privacy-by-design should be organized across the CCR lifecycle. As shown in Figure 2, the synthesis supports a lifecycle model in which protection is not confined to a final verification phase. Instead, security and privacy controls should be considered across data collection and sensing, creative data and model use, co-creative interaction, embodiment and planning, testing and deployment, and governance and accountability.

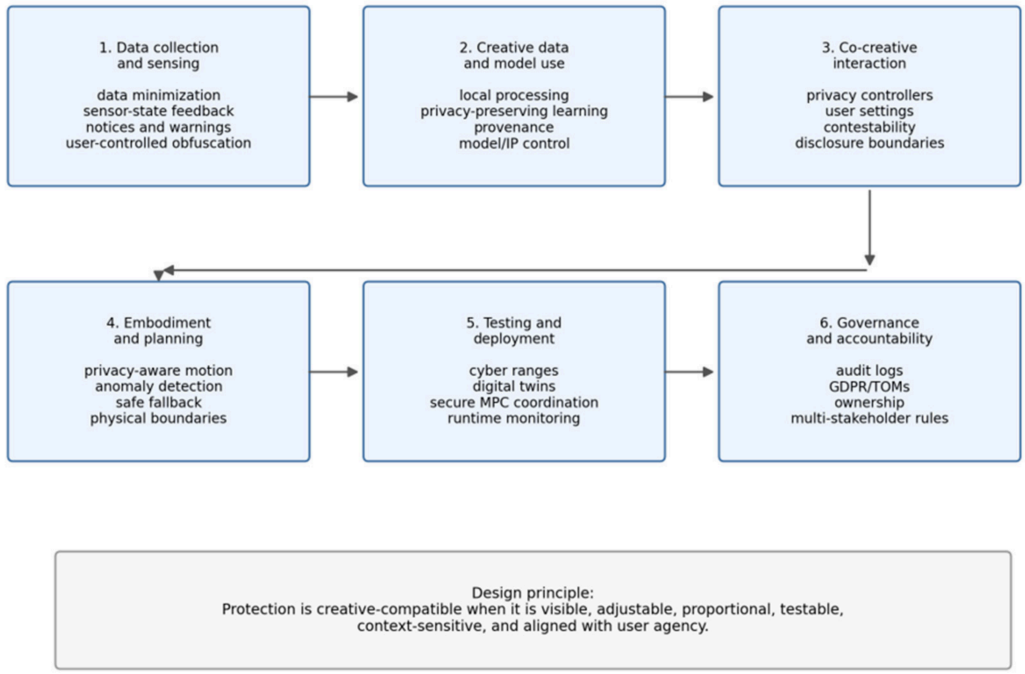


Figure 2. Lifecycle-oriented integration model for cybersecurity-by-design and privacy-by-design in Co-Creative Robotics.

The model presented in Figure 2 is a conceptual synthesis of the lifecycle requirements emerging from the reviewed evidence and is intended to guide future CCR research and design, rather than to represent an empirically validated framework. It should be read as an iterative lifecycle rather than as a rigid linear pipeline. Decisions made in early phases, such as data minimization or sensor-state feedback, affect what can be trained, remembered, disclosed, tested, and governed later. Conversely, deployment evidence, runtime monitoring, incidents, and user feedback should inform revisions to sensing policies, interaction rules, model governance, and accountability mechanisms.

Table 10 maps the lifecycle phases represented in Figure 2 to the main controls and evidence emerging from the corpus. The table shows that future work should not focus only on isolated privacy or security mechanisms, but on how those mechanisms are combined and made visible, adjustable, testable, and compatible with creative agency.

Table 10. Lifecycle phases, controls and implications for secure and privacy-aware CCR.

Lifecycle Phase	Main Controls or Design Focus	Implication for CCR
Data collection and sensing	Data minimization, local processing where possible, accessible privacy notices, sensor-state communication, user-controlled obfuscation, edge-based background blurring or selective capture, privacy-respecting embodied behaviors, and policies for minimizing bystander data [11,21,23,51].	Primary users, collaborators, and bystanders should understand what the robot senses, while the system should preserve sensing-based collaboration and limit the capture, transmission, and retention of ambient data from non-participating persons.
Creative data and model use	Privacy-preserving learning, dataset governance, adversarial robustness, provenance, and model or intellectual-property controls [43,45,47].	Creative traces, drafts, prompts, preferences, and co-created outputs should be treated as sensitive creative assets rather than ordinary telemetry.
Co-creative interaction	Privacy controllers, context-sensitive disclosure, user-adjustable settings, contestability, notifications, and meaningful consent mechanisms [17,18,22,27].	Protection should be embedded in the interaction so that users can negotiate, contest, or redirect sensing, memory, disclosure, and robotic contribution.
Embodiment and planning	Privacy-aware motion planning, independent anomaly detection, safe fallback, spatial boundaries, and privacy-preserving multi-robot coordination [6,15,24].	Security and privacy should account for where the robot moves, how it points its sensors, how it acts physically, and how it coordinates with other robots.
Testing and deployment	Cyber ranges, digital twins, runtime monitoring, controlled attack testing, and continuous improvement [7,8].	Open-ended creative behavior should be validated under controlled conditions before and during real-world deployment.
Governance and accountability	Audit logs, documentation, GDPR-oriented technical and organizational measures, ownership rules, accountability mechanisms, participatory modeling of stakeholders and consent roles, and multi-stakeholder governance [4,10,12,13,26,28].	Responsibility for privacy, security, creative ownership, consent, bystander protection, and robot-generated or co-produced outputs should be traceable and reviewable across the system lifecycle.

To translate the lifecycle synthesis into an actionable developer-oriented checklist, a CCR team should verify before deployment and during operation that: (1) the scope of sensing, bystander handling, and sensor-state communication are explicitly defined; (2) creative traces, prompts, preferences, and co-produced outputs have documented retention, reuse, deletion, and provenance rules; (3) users can inspect, adjust, contest, or override privacy settings and robotic contributions; (4) motion and actuation are protected through context-sensitive anomaly detection and safe fallback mechanisms; (5) validation scenarios test both attacks and legitimate expressive deviations, so that protection mechanisms do not suppress creative behavior; and (6) consent states, configuration changes, incidents, and accountability responsibilities remain auditable across the lifecycle.

The following discussion elaborates on these checks across the lifecycle phases shown in Figure 2 and Table 10, illustrating how they can be operationalized in the design, deployment, and governance of CCR systems.

During data collection and sensing, CCR systems should apply data minimization, local processing where possible, accessible privacy notices, user-controlled obfuscation, selective capture, visible sensor-state communication, and embodied behaviors that make privacy protection perceptible [11,21,23,51]. This phase is foundational because co-creative interaction often begins with multimodal capture of speech, images, gestures, gaze, proximity, unfinished artifacts, and behavioral cues.

For ambient privacy management, participatory modeling should extend the lifecycle analysis beyond the primary user by mapping the roles, rights, exposure levels, and consent expectations of primary users, co-present collaborators, bystanders, operators, and other affected stakeholders. In a shared design studio or makerspace, for example, a CCR system may need camera-based tracking of the active user's hands, materials, or workspace to provide creative scaffolding, while avoiding unnecessary capture of other people working nearby. This can be operationalized through user-selectable regions of interest, real-time obfuscation of background areas, edge-based processing before storage or transmission, selective audio activation, spatial no-sensing zones, and visible sensor-state indicators. In this way, ambient privacy controls can be integrated into the sensing lifecycle without disabling the contextual information needed for co-creative collaboration.

During creative data and models use, privacy-preserving learning, dataset governance, adversarial robustness, provenance, and model or intellectual property controls become relevant [43,45,47]. In CCR, this phase should also account for the status of sketches, prompts, design traces, drafts, preferences, and co-created outputs as potentially sensitive creative assets. Protection should cover both personal data and creative material that may reveal the user's intentions, style, identity, or unfinished ideas.

During co-creative interaction, privacy should be managed as an interactional property. Robots should support conversational privacy boundaries, context-sensitive disclosure, user-adjustable settings, contestability, and meaningful consent [17,18,22,27]. In family, educational, healthcare, or care contexts, controls should also support multi-stakeholder negotiation rather than assuming a single user with full authority over all data [25,28].

During embodiment and planning, robots require cyber-physical safeguards that account for physical actuation, sensing direction, spatial access, and motion choices. Anomaly detection, independent sensing, privacy-aware motion planning, and safe fallback mechanisms are important because creative robots may operate in open environments where rigid scripts are not sufficient [6,15]. During multi-robot or distributed operation, secure multi-party computation and related privacy-preserving coordination techniques can support collaboration without exposing private valuations, tasks, or preferences [24].

During testing and deployment, cyber ranges, digital twins, and runtime monitoring can support safe experimentation, attack testing, countermeasure assessment, and continuous improvement [7,8]. This phase is especially important for CCR because creative autonomy may produce variable behaviors that cannot be exhaustively specified at design time. Controlled validation can enable more ambitious interactions while reducing risks to users and the environment.

During governance and accountability, provenance, audit logs, documentation, accountability mechanisms, ownership rules, and domain-specific compliance should address creative ownership, privacy, consent, bystander protection, and responsibility for robot-generated or co-produced outputs. For shared CCR settings, governance should also specify who can configure sensing regions, activate or override obfuscation, approve retention of creative traces, review captured data, and request deletion or exclusion of ambient data.

GDPR-oriented technical and organizational measures, together with multi-stakeholder governance, can help make these responsibilities explicit and auditable [4,10,12,13,26,28].

Answer to SRQ5. The strongest implication is that future CCR research should move from isolated controls toward lifecycle architectures that connect sensing, ambient privacy management, model use, co-creative interaction, embodiment, deployment, monitoring, and governance. The remaining uncertainty concerns how responsibilities should be distributed among users, designers, operators, institutions, and affected stakeholders as decisions regarding sensing, obfuscation, retention, consent, deletion, and accountability evolve over time. A key research gap is the empirical validation of such lifecycle models in shared CCR environments where protection must safeguard co-present collaborators and bystanders without eroding the primary user's sensing-based creative experience.

4.8. Answering the Main Research Question

The synthesis supports a conditional answer to the main research question. Cybersecurity and privacy can be integrated into CCR without necessarily constraining creativity and autonomy, but this depends on how protection is designed. If protection is added as a rigid or opaque constraint, it risks limiting exploration, improvisation, and user agency. If it is designed as a visible, adjustable, and context-sensitive part of the interaction, it can instead support trust and make creative collaboration more sustainable.

The evidence indicates that the key issue is not the presence of controls, but their fit with the creative and embodied nature of CCR. Co-creative robots need room for adaptation, ambiguity, and initiative, but they also require clear boundaries around sensing, memory, disclosure, motion, data reuse, and accountability. Protection is compatible with CCR when it helps users understand what the robot is doing, retain meaningful control over the interaction, and contest or redirect robot behavior when needed.

More concretely, security and privacy controls may become constraining when they disrupt the temporal, contextual, or embodied continuity of the co-creative process. Encryption, remote verification, or privacy-preserving computation may introduce latency if they are not designed for real-time interaction, thereby interrupting turn-taking, improvisation, or the user's flow state. Strict token or memory limits may prevent the robot from recalling earlier sketches, gestures, preferences, or design alternatives, reducing its ability to scaffold the creative process over time. Similarly, overly rigid data minimization may remove contextual cues such as gaze, posture, spatial position, or affective signals that are needed for situated creative support. Finally, overly restrictive security filters or anomaly thresholds may classify unusual yet expressive robot movements or user inputs as unsafe, even when such deviations are part of exploration, play, or artistic experimentation. These examples clarify that the problem is not protection itself, but protection whose timing, granularity, or flexibility is misaligned with co-creative interaction.

This leads to three integration requirements.

- First, protection should be participatory, because users and affected stakeholders should contribute to defining relevant assets, harms, and acceptable trade-offs.
- Second, it should be creative-compatible, because security and privacy mechanisms should preserve exploration, interpretive flexibility, and human initiative.
- Third, it should be lifecycle-aware, because the risks of CCR change across sensing, model use, interaction, embodiment, deployment, monitoring, and governance.

Main RQ answer. Cybersecurity and privacy can be integrated into Co-Creative Robotics without constraining creativity and autonomy when they are implemented as participatory, transparent, adjustable, context-sensitive, and lifecycle-aware mechanisms that support trust and user agency. They become constraining when they are opaque, rigid, disproportionate, mechanically transferred from other domains, or misaligned with the

temporal, contextual, and embodied continuity of the creative interaction they are meant to protect.

5. Discussion

The results lead to three cross-cutting implications for CCR design and evaluation. These implications do not restate the individual SRQ answers; rather, they connect them to the broader design problem of making protection compatible with creative agency.

Protection as Creative-Compatible Infrastructure: Cybersecurity-by-design and privacy-by-design should be treated as part of CCR's creative infrastructure. Co-creative robots rely on sensing, memory, adaptation, suggestion, movement, and sometimes distributed coordination. These capabilities can support creativity, but they also introduce boundaries that users need to understand and influence. Protection is most compatible with CCR when it clarifies those boundaries rather than merely restricting the system after risks have already emerged.

This requires a shift from generic control to situated control. In CCR, the relevant design question is not only whether data are encrypted or access is restricted, but also whether users can understand what the robot senses, how it uses creative traces, when information may be retained or disclosed, and how robotic action can be redirected. This framing makes cybersecurity and privacy part of the interaction design problem rather than a separate compliance layer.

From Isolated Mechanisms to Lifecycle Integration: The reviewed evidence offers many useful mechanisms, including privacy settings, sensor-state communication, disclosure control, privacy-aware behavior, privacy-aware motion, anomaly detection, testbeds, digital twins, and threat modeling. These mechanisms are insufficient if applied in isolation. A notice does not address cyber-physical risk; anomaly detection does not address consent or disclosure; and threat modeling does not preserve creative agency unless it is connected to how users actually collaborate with the robot.

The lifecycle proposed in this paper addresses this problem by connecting protection to data collection and ambient privacy management, model use, co-creative interaction, embodiment, deployment, monitoring, and governance. This does not imply that every CCR system requires every possible control. Rather, it means that design teams should decide which controls are needed at each phase and should justify how those controls preserve user agency, creative flexibility, bystander protection, and accountability.

Implications for Evaluation: The synthesis also suggests that evaluation criteria for CCR should become multidimensional. A system may be technically secure but creatively sterile, or creatively engaging but privacy-invasive. Future evaluations should consider creativity, autonomy, privacy, including bystander protection, security, trust, and usability together. In particular, privacy and security mechanisms should be assessed not only for technical effectiveness, but also for their impact on exploration, self-disclosure, perceived control, willingness to collaborate, and the user's ability to contest or redirect the robot.

To make this evaluation agenda more operational, Table 11 summarizes potential evaluation dimensions and indicators for use in future CCR studies. The table is not intended as a fixed set of metrics, but as a starting point for designing evaluations that jointly consider protection, creativity, autonomy, and interaction quality.

These dimensions should be evaluated jointly because a CCR system may be technically secure but creatively disruptive or creatively engaging but privacy-invasive. This is especially important because several privacy mechanisms affect behavior in non-obvious ways: warnings, notifications, settings, and explanations may improve awareness or acceptance, but they do not automatically produce meaningful consent, reduced disclosure, or greater perceived safety. Therefore, future CCR evaluations should report not only the

technical effectiveness of protection mechanisms, but also their impact on creative agency, interaction continuity, user control, trust, and actual user behavior, especially in realistic, embodied settings.

Table 11. Possible evaluation dimensions for secure and privacy-aware Co-Creative Robotics.

Evaluation Dimension	What Should Be Assessed	Possible Indicators or Metrics
Creativity preservation	Whether security and privacy controls preserve the user's ability to explore, improvise, and generate alternatives.	Novelty, diversity, and fluency of ideas; perceived creative agency; perceived disruption of creative flow; number of user-initiated alternatives.
Autonomy and agency	Whether users can understand, contest, redirect or override robot actions and data practices.	Perceived control, successful override or contest actions, clarity of robot explanations, and ability to modify privacy or interaction settings.
Privacy protection	Whether primary users, co-present collaborators, and bystanders understand and control sensing, disclosure, obfuscation, retention, and reuse of data.	Sensing awareness; consent comprehension; disclosure behavior; bystander data minimization; use of regions of interest, obfuscation, privacy zones, or edge-based processing; perceived privacy.
Cybersecurity and safety	Whether the system detects, resists and recovers from cyber-physical threats without excessive false alarms.	Anomaly detection accuracy; false positive and false negative rates; recovery time; safe fallback activation; attack-resilience testing.
Interaction quality	Whether protection mechanisms preserve fluid, timely and understandable co-creative interaction.	Latency; interruption frequency; task completion; perceived usability; trust; willingness to collaborate; conversational or gestural continuity.
Governance and accountability	Whether decisions, data flows, consent states, bystander protections, and creative outputs remain traceable and reviewable.	Auditability; consent traceability; provenance of co-created outputs; documented responsibility; roles for configuring sensing or obfuscation; availability of review, deletion or ambient-data exclusion mechanisms.

The mixed evidence on transparency-oriented privacy mechanisms also requires a cautious interpretation. Across the reviewed studies, warnings, notifications, and configurable privacy settings often improved awareness, comprehension, acceptance, or perceived control, but their effects on actual disclosure behavior, perceived safety, and sustained user decisions were not consistent [20,22,27]. For CCR, this distinction is particularly important because a mechanism may be understandable yet behaviorally ineffective, and still impose cognitive or interactional burdens during an open-ended creative activity. Future evaluations should therefore distinguish informational outcomes, such as awareness and comprehension, from behavioral and experiential outcomes, such as actual disclosure, meaningful consent, perceived safety, creative continuity, and willingness to collaborate over repeated embodied interactions.

Limitations and Future Research

This review has limitations. First, CCR is still an emerging and inconsistently labeled field, so the search strategy cannot guarantee complete coverage. Second, only a subset of the final corpus explicitly addresses CCR or creative robotics. Many studies were included because their mechanisms are transferable from social robotics, domestic robotics, healthcare robotics, multi-robot systems, cyber-physical security, or privacy engineering. Third, the protocol was not preregistered, although the selection, eligibility, quality assessment, and extraction procedures were documented in structured records. Fourth, excluding preprints, viewpoints, reviews, theses, and short papers improved evidential consistency but may have left out early-stage ideas. Fifth, the synthesis is qualitative because the included studies differ substantially in methods, domains, and types of evidence.

The 2020–2026 publication window also limits the temporal breadth of the review. Although it prioritizes recent developments, it may omit earlier work in robotics, HRI, privacy, and security that could inform CCR, while the 2026 literature was necessarily incomplete at the time of the search. Moreover, restricting the primary corpus to peer-reviewed, English-language, and accessible full-text publications may introduce publication and availability bias by favoring more mature or positively reported results. Consequently, the proposed lifecycle agenda should be interpreted as a synthesis of the recent accessible evidence rather than as a complete or universally generalized account of secure and privacy-aware CCR.

A further limitation concerns the conceptual framing of Co-Creative Robotics. Reference [1], which contributes to the definition and positioning of CCR used in this review, is authored by members overlapping with the current author team. References [33,35,37] also include overlapping authors and were used as related/background work. This overlap is partly a consequence of the field's early stage, in which a small number of contributions have helped define the terminology and research agenda. Nevertheless, it may have influenced how the review frames CCR, particularly in its emphasis on co-creation, creative autonomy, and human–robot collaboration. To mitigate this risk, the review distinguished primary studies from background references, applied explicit eligibility and quality-assessment criteria, used multi-author screening and consensus procedures, and incorporated evidence from adjacent domains such as HRI, social-robot privacy, cyber-physical security, and autonomous systems. Future reviews should revisit this framing as the CCR literature broadens and becomes more independent.

Future research should move from transferability to direct CCR validation. This includes developing CCR-specific security and privacy patterns, testing them with physical co-creative robots, and evaluating their effect on both protection and creativity. Longitudinal studies are also needed because privacy expectations, trust, agency, and creative dependence may change as robots learn user preferences, retain creative histories, and become embedded in everyday routines. Finally, participatory governance should involve creators, educators, workers, families, patients, children, and other affected stakeholders in deciding which creative assets require protection and how autonomy should be bounded.

6. Conclusions

This PRISMA-informed review examined whether cybersecurity-by-design and privacy-by-design can be integrated into Co-Creative Robotics without constraining creativity and autonomy. The final corpus of 27 primary studies supports a conditional answer. Protection can support CCR when it is designed as an adaptive and intelligible part of the interaction. It becomes constraining when it is opaque, disproportionate, rigid, or introduced only after creative interaction patterns have already been fixed.

However, much of the available evidence still comes from adjacent robotic and human–robot interaction domains, and the effects of transparency-oriented privacy mechanisms remain mixed: increased awareness or acceptance does not consistently translate into reduced disclosure or greater perceived safety.

The review presents a synthesis connecting creative-HRI evidence, social-robot privacy studies, cyber-physical security, security modeling, and responsible robotics. Its main contribution is the lifecycle interpretation of secure and privacy-aware CCR: protection should be considered across sensing, data, and model use, co-creative interaction, embodiment, deployment, monitoring, and governance. The resulting lifecycle model should be understood as a conceptual synthesis and research agenda whose effectiveness still requires empirical validation in realistic CCR environments.

Across these phases, the goal is not to remove all constraints, but to make constraints understandable, adjustable, testable, and aligned with user agency.

The resulting agenda is practical. CCR systems should make sensing and disclosure visible, support meaningful user control, protect creative traces and co-produced outputs, test embodied behavior before deployment, monitor risks over time, and document responsibility for autonomous or co-created actions. In this framing, cybersecurity and privacy are not external limits on creativity. They are conditions for trustworthy, accountable, and sustainable co-creative autonomy.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info17080771/s1, Table S1: PRISMA 2020 checklist.

Author Contributions: Conceptualization, E.M.; methodology, E.M. and S.N.-H.; formal analysis, E.M., S.N.-H., A.C. and J.B.; investigation, E.M., S.N.-H., A.C. and J.B.; data curation, E.M., S.N.-H., A.C. and J.B.; writing—original draft preparation, E.M.; writing—review and editing, E.M., S.N.-H., A.C. and J.B.; visualization, E.M., S.N.-H., A.C. and J.B.; supervision, E.M. and S.N.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported and funded by the project CYBER–Progetto integrato 2.1 Cybersecurity dei sistemi energetici per la transizione energetica-digitale–MASE 2026, EU-funded project CAIOC with Grant Agreement ID: 101190339, and Portugal national funds through the FCT—Foundation for Science and Technology, IP, within the scope of the project UID/00066/2025 with the DOI identifier https://doi.org/10.54499/UID/00066/2025.

Institutional Review Board Statement: Not applicable. This study is a literature review and did not involve new studies with humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors used generative AI and AI-assisted tools to support language refinement, formatting, and presentation. All scientific content, analyses, interpretations, and conclusions were developed, reviewed, and validated by the authors, who take full responsibility for the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nikghadam-Hojjati, S.; Marchetti, E.; Barata, J.; Calabrò, A. Moving into Co-Creative Robotics. In Proceedings of the 20th International Conference on Web Information Systems and Technologies (WEBIST 2024), Porto, Portugal, 17–19 November 2024; pp. 307–314. [CrossRef]
2. Ali, S.; Devasia, N.; Park, H.W.; Breazeal, C. Social Robots as Creativity Eliciting Agents. *Front. Robot. AI* **2021**, *8*, 673730. [CrossRef] [PubMed]

3. Song, B.; Soria Zurita, N.F.; Nolte, H.; Singh, H.; Cagan, J.; McComb, C. When Faced With Increasing Complexity: The Effectiveness of Artificial Intelligence Assistance for Drone Design. *J. Mech. Des.* **2022**, *144*, 021701. [CrossRef]
4. European Parliament; Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC. *Off. J. Eur. Union* **2016**, *L119*, 1–88.
5. Radanliev, P. Digital Security by Design. *Secur. J.* **2024**, *37*, 1640–1679. [CrossRef]
6. Kayan, H.; Heartfield, R.; Rana, O.; Burnap, P.; Perera, C. CASPER: Context-Aware IoT Anomaly Detection System for Industrial Robotic Arms. *ACM Trans. Internet Things* **2024**, *5*, 18. [CrossRef]
7. Roberts, A.; Maennel, O.; Snetkov, N. Cybersecurity Test Range for Autonomous Vehicle Shuttles. In Proceedings of the 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Vienna, Austria, 6–10 September 2021; pp. 239–248. [CrossRef]
8. Shaikh, E.; Al-Ali, A.R.; Muhammad, S.; Mohammad, N.; Aloul, F. Security Analysis of a Digital Twin Framework Using Probabilistic Model Checking. *IEEE Access* **2023**, *11*, 26358–26374. [CrossRef]
9. Thompson, R.E.; McLaughlin, M.; Powers, C.; Votipka, D. There Are Rabbit Holes I Want to Go Down That I'm Not Allowed to Go Down: An Investigation of Security Expert Threat Modeling Practices for Medical Devices. In Proceedings of the 33rd USENIX Security Symposium, Philadelphia, PA, USA, 14–16 August 2024; pp. 4909–4926.
10. Callari, T.C.; Vecellio Segate, R.; Hubbard, E.-M.; Daly, A.; Lohse, N. An Ethical Framework for Human-Robot Collaboration for the Future People-Centric Manufacturing: A Collaborative Endeavour with European Subject-Matter Experts in Ethics. *Technol. Soc.* **2024**, *78*, 102680. [CrossRef]
11. Zhang, L.; Stangl, A.; Sharma, T.; Tseng, Y.-Y.; Xu, I.; Gurari, D.; Wang, Y.; Findlater, L. Designing Accessible Obfuscation Support for Blind Individuals' Visual Privacy Management. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2024), Honolulu, HI, USA, 11–16 May 2024; 19 p. [CrossRef]
12. Ghoshal, A.; Brandão, M.; Abu-Salma, R.; Modgil, S. Embodied AI at the Margins: Postcolonial Ethics for Intelligent Robotic Systems. *Proc. AAAI/ACM Conf. AI Ethics Soc.* **2025**, *8*, 1120–1133. [CrossRef]
13. Trentesaux, D.; Karnouskos, S. Ethical Behaviour Aspects of Autonomous Intelligent Cyber-Physical Systems. In *Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future*; Borangiu, T., Trentesaux, D., Thomas, A., Leitão, P., Oliveira, J., Eds.; Springer: Cham, Switzerland, 2020; pp. 55–71. [CrossRef]
14. Wohlin, C. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE 2014), London, UK, 13–14 May 2014; p. 38. [CrossRef]
15. Shome, R.; Kingston, Z.; Kavraki, L.E. Robots as AI Double Agents: Privacy in Motion Planning. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 2861–2868. [CrossRef]
16. Lutz, C.; Tamò-Larrieux, A. The Robot Privacy Paradox: Understanding How Privacy Concerns Shape Intentions to Use Social Robots. *Hum. Mach. Commun.* **2020**, *1*, 87–111. [CrossRef]
17. Dietrich, M.; Krüger, M.; Weisswange, T.H. What Should a Robot Disclose About Me? A Study About Privacy-Appropriate Behaviors for Social Robots. *Front. Robot. AI* **2023**, *10*, 1236733. [CrossRef] [PubMed]
18. Tang, B.; Sullivan, D.; Cagiltay, B.; Chandrasekaran, V.; Fawaz, K.; Mutlu, B. CONFIDANT: A Privacy Controller for Social Robots. In Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2022), Sapporo, Japan, 7–10 March 2022; pp. 205–214.
19. Lutz, C.; Tamò-Larrieux, A. Do Privacy Concerns About Social Robots Affect Use Intentions? Evidence From an Experimental Vignette Study. *Front. Robot. AI* **2021**, *8*, 627958. [CrossRef] [PubMed]
20. Aryania, A.; Huertas-Garcia, R.; Forgas-Coll, S.; Angulo, C.; Alenyà, G. Effects of Privacy Warning on the Intention to Disclose Personal Information During Interaction with a Robot in Public Spaces. *Int. J. Soc. Robot.* **2025**, *17*, 1055–1073. [CrossRef]
21. Yang, D.; Chae, Y.-J.; Kim, D.; Lim, Y.; Kim, D.H.; Kim, C.; Park, S.-K.; Nam, C. Effects of Social Behaviors of Robots in Privacy-Sensitive Situations. *Int. J. Soc. Robot.* **2022**, *14*, 589–602. [CrossRef]
22. Stapels, J.G.; Penner, A.; Diekmann, N.; Eyssel, F. Never Trust Anything That Can Think for Itself, if You Can't Control Its Privacy Settings: The Influence of a Robot's Privacy Settings on Users' Attitudes and Willingness to Self-disclose. *Int. J. Soc. Robot.* **2023**, *15*, 1487–1505. [CrossRef]
23. Windl, M.; Leusmann, J.; Schmidt, A.; Feger, S.S.; Mayer, S. Privacy Communication Patterns for Domestic Robots. In Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS 2024), Philadelphia, PA, USA, 12–13 August 2024; pp. 121–139.
24. Alsayegh, M.; Vanegas, P.; Newaz, A.A.R.; Bobadilla, L.; Shell, D.A. Privacy-Preserving Multi-Robot Task Allocation via Secure Multi-Party Computation. In Proceedings of the 2022 European Control Conference (ECC), London, UK, 12–15 July 2022; pp. 1274–1281. [CrossRef]

25. Jayaraman, S.; Phillips, E.K.; Church, D.; Riek, L.D. Privacy and Utility Perceptions of Social Robots in Healthcare. *Comput. Hum. Behav. Artif. Hum.* **2024**, *2*, 100039. [CrossRef]
26. Horstmann, B.; Diekmann, N.; Buschmeier, H.; Hassan, T. Towards Designing Privacy-Compliant Social Robots for Use in Private Households: A Use Case Based Identification of Privacy Implications and Potential Technical Measures for Mitigation. In Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Virtual Conference, 31 August–4 September 2020; pp. 869–876. [CrossRef]
27. Sakai, K.; Mitsuno, S.; Ban, M.; Yoshikawa, Y.; Shimpo, F.; Harata, S.; Ishiguro, H. Effect of Robot Notification on Acquiring Permission to Use Personal Information. In Proceedings of the International Conference on Human-Agent Interaction (HAI 2023), Gothenburg, Sweden, 4–7 December 2023; pp. 124–132. [CrossRef]
28. Levinson, L.; McKinney, J.; Nippert-Eng, C.; Gomez, R.; Šabanović, S. Our Business, Not the Robot's: Family Conversations About Privacy with Social Robots in the Home. *Front. Robot. AI* **2024**, *11*, 1331347. [CrossRef] [PubMed]
29. Boden, M.A. *The Creative Mind: Myths and Mechanisms*, 2nd ed.; Routledge: London, UK, 2004.
30. Wiggins, G.A. A Preliminary Framework for Description, Analysis and Comparison of Creative Systems. *Knowl.-Based Syst.* **2006**, *19*, 449–458. [CrossRef]
31. Shneiderman, B. Creativity Support Tools: Accelerating Discovery and Innovation. *Commun. ACM* **2007**, *50*, 20–32. [CrossRef]
32. Babaei, Z.; Kalateh, S.; Nikghadam-Hojjati, S.; Leitão, P.; Barata, J. Human Robot Collaborative Creativity: Challenges and Opportunities. In *Technological Innovation to Tackle Societal Challenges*; IFIP Advances in Information and Communication Technology; Camarinha-Matos, L.M., Ferrada, F., Eds.; Springer: Cham, Switzerland, 2026; Volume 790, pp. 313–328. [CrossRef]
33. Marchetti, E.; Costa, A.; Nikghadam-Hojjati, S.; Barata, J.; Calabrò, A. Toward Cybersecure-by-Design Manufacturing Systems: A Taxonomy of Hardware-Based Approaches for Software Cybersecurity Violation Detection. In *Innovative Intelligent Industrial Production and Logistics*; Communications in Computer and Information Science; Barata, J., Madani, K., Panetto, H., Eds.; Springer: Cham, Switzerland, 2026; Volume 2825, pp. 372–392. [CrossRef]
34. Ross, R.; McEvilley, M.; Carrier Oren, J. *Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems*; NIST Special Publication 800-160, Volume 1, Revision 1; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2022. [CrossRef]
35. Calabrò, A.; Marchetti, E.; Nikghadam-Hojjati, S. ACCURATE-eternAl infrastruCture for seCUrity in softwaRe and hArdware developmenT and assessmEnt. In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025), Marbella, Spain*; SCITEPRESS: Setúbal, Portugal, 2025; Volume 1, pp. 157–164. [CrossRef]
36. Cavoukian, A. *Privacy by Design: The 7 Foundational Principles*; Information and Privacy Commissioner of Ontario: Toronto, ON, Canada, 2011.
37. Silva, A.d.N.; Nikghadam-Hojjati, S.; Estrada Jimenez, L.A.; Barata, J. Behavioral Dynamics of Creative Cooperation in Human-3C-Bot Communities: An Agent-Based Simulation of Trust-Mediated Innovation. *IEEE Access* **2026**, *14*, 7907–7920. [CrossRef]
38. High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
39. IEEE. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, 1st ed.; IEEE: Piscataway, NJ, USA, 2019.
40. Shostack, A. *Threat Modeling: Designing for Security*; Wiley: Indianapolis, IN, USA, 2014.
41. Deng, M.; Wuyts, K.; Scandariato, R.; Preneel, B.; Joosen, W. A Privacy Threat Analysis Framework: Supporting the Elicitation and Fulfillment of Privacy Requirements. *Requir. Eng.* **2011**, *16*, 3–32. [CrossRef]
42. Spencer, M.; Coles-Kemp, L.; Hansen, R.R. Navigating the Landscape of Security Modelling: The MORS Grid. *J. Cybersecur.* **2024**, *10*, tyae024. [CrossRef]
43. Arora, S.; Hastings, J. Securing Agentic AI Systems: A Multilayer Security Framework. In *Proceedings of the 2025 5th International Conference on Robotics, Automation, and Artificial Intelligence (RAAI)*; IEEE: Singapore, 2025; pp. 509–514. [CrossRef]
44. Rousi, R.; Mäkitalo, N.; Samani, H.; Kemell, K.-K.; de Cerqueira, J.S.; Vakkuri, V.; Mikkonen, T.; Abrahamsson, P. GPT versus Humans: Uncovering Ethical Concerns in Conversational Generative AI-Empowered Multi-Robot Systems. *arXiv* **2024**, arXiv:2411.14009.
45. Al-Mahmood, A.; Al-Hakeem, S.M.; Mahmood, M.T.; Hussein, A.M.; Kurdi, I.A.; Ismael, A.A.; Fadel, H.A.; Al-Janabi, A.F.H.; Algburi, S. Blockchain for Intellectual Property Protection: Securing Digital Content Ownership and Rights Management. In Proceedings of the 2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 1–2 May 2025. [CrossRef]
46. Garg, A.; Verma, N.; Singhal, R.K.; Sharma, H.; Sharma, H.; Balyan, P. Decentralized Multi-Robot Systems Using Blockchain and New Distributed Ledger Technologies. In Proceedings of the 2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 28–30 November 2025. [CrossRef]
47. Yang, Q.; Huang, A.; Fan, L.; Chan, C.S.; Lim, J.H.; Ng, K.W.; Ong, D.S.; Li, B. Federated Learning with Privacy-Preserving and Model IP-Right-Protection. *Mach. Intell. Res.* **2023**, *20*, 19–37. [CrossRef]

48. Beatrys, B.; Peruzzo, A. Real Harassment, Virtual Robots? Exploring Misogyny Against Machines. *Int. J. Commun.* **2025**, *19*, 3090–3110.
49. Giansanti, D.; Gulino, R.A. The Cybersecurity and the Care Robots: A Viewpoint on the Open Problems and the Perspectives. *Healthcare* **2021**, *9*, 1653. [CrossRef] [PubMed]
50. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
51. De Coninck, S.; Wang, W.-C.; Leroux, S.; Simoens, P. Privacy-Preserving Visual Analysis: Training Video Obfuscation Models Without Sensitive Labels. *Appl. Intell.* **2024**, *54*, 6041–6052. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.