

Cognitive Computing Frameworks for Adaptive Human-Robot Collaboration

Subramaniam S^{1*}, Ravichandran N² and Venkatesan S³

¹Department of Robotics and Automation, Noyyal Center for Future Computing Architectures, Bangladesh Email: sathish.subramaniam@noyyal.edu

ORCID: <https://orcid.org/3030-7495-9223-8659>

²Department of Artificial Intelligence and Data Science, Meenachil Institute of Distributed Intelligence, Taiwan Email: nandini.ravichandran@meenachil.edu

ORCID: <https://orcid.org/4606-1851-3773-2172>

³Department of Artificial Intelligence and Data Science, Achankovil Center for Applied Machine Learning, Taiwan Email: shalini.venkatesan@achankovil.edu

ORCID: <https://orcid.org/1624-9208-6402-0161>

ABSTRACT

Collaborative robots increasingly share a workspace, a task, and even a single object with human partners, a shift that requires software capable of representing what a human collaborator is doing, intending, and about to need rather than treating the human as just another obstacle. This article synthesizes the literature on the cognitive computing frameworks built to meet that requirement: symbolic cognitive architectures such as Soar and ACT-R, and their embodiment in physical robots through extensions like ACT-R/E, that supply a representational backbone for robot reasoning; perception and intention-recognition pipelines that have moved from rule-based heuristics to deep sequence models reporting high accuracy on constrained benchmarks; planning and decision-making layers that treat human trust, intention, and preference as evolving hidden state rather than fixed input, often within formal frameworks such as partially observable Markov decision processes; and the reinforcement-learning, learning-from-demonstration, and digital-twin methods that let all of the above keep adapting after deployment. Read together, the evidence supports cognitive computing as a coherent and increasingly well-evidenced architectural pattern for adaptive human-robot collaboration, while leaving evaluation at realistic deployment scale, two-way legibility between robot and human, and integration of symbolic and learned reasoning as a substantially open research agenda.

Keywords: cognitive computing; human-robot collaboration; cognitive architecture; intention recognition; trust-aware planning

Citation: Subramaniam S, Ravichandran N and Venkatesan S [2022]. Cognitive Computing Frameworks for Adaptive Human-Robot Collaboration. DOI: <https://doi.org/10.5281/zenodo.22843848>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 8 January 2022 Accepted: 29 March 2022 Published: 7 July 2022

Article Class: Research Article

1. Introduction

A cobot arm that hands a technician the right fastener before she asks for it, a warehouse robot that yields a narrow aisle because it has inferred a human is about to reach across it, a surgical assistant that anticipates the next instrument in a sequence it has only seen once — none of these behaviors come from faster motors or better grippers. They come from software that represents what the human partner is likely doing, intending, and about to need, and that updates that representation continuously as the collaboration unfolds. That is, in a fairly literal sense, what "cognitive computing" was coined to describe: systems that reason, learn, and adapt over unstructured, uncertain input in a way that mimics some of the flexibility of human cognition, as opposed to executing a fixed, pre-specified control loop [1]. Robotics has always needed perception and control; what human-robot collaboration specifically needs, and what a decade of published work increasingly delivers, is a cognitive layer sitting between the two — one that models the human, not just the workspace.

Goodrich and Schultz's early, still widely cited survey of human-robot interaction framed the field's central design questions — the level and mode of a robot's autonomy, how information should flow between human and robot, and how a team's performance should even be measured — well before "cognitive computing" was a term anyone applied to robotics, and much of what follows in this article can be read as later, more technically specific answers to the questions their survey first laid out [2]. This is a genuinely different design problem than industrial robotics solved for most of its history. A robot welding a car chassis on a fixed line does not need a model of the humans around it, because by design there are not supposed to be any within its operating envelope; the entire safety architecture, historically, was about physical separation rather than mutual awareness. Collaborative robots — cobots — invert that premise, sharing a workspace, a task, and often a single object with a human partner in real time, and the moment that premise is inverted, the robot needs something it never needed before: an internal, updating estimate of what the human is doing and is about to do. Villani et al.'s widely cited survey of human-robot collaboration in industrial settings frames this shift explicitly around three technical pillars — safety, intuitive interaction, and applications — and

argues that none of the three is solvable by mechanical design alone; each depends on the robot maintaining some model of the human that goes beyond a point cloud of obstacle positions [3]. Singh, Rantanen, and Heard's recent comprehensive survey of human-robot teaming organizes almost the entire span of this literature around exactly three interlocking pillars -- collaboration, communication, and cognition -- and argues, consistent with the framing adopted throughout this article, that progress on any one pillar in isolation tends to stall without corresponding progress on the other two: a robot with a rich cognitive model of its partner but no channel to communicate through that model is not meaningfully more collaborative than one without it [4]. Cognitive computing frameworks are the software answer to that dependency, and this article surveys what the literature has built along that line: architectures that give a robot something like working memory and declarative knowledge about a task, perception pipelines that turn raw sensor streams into inferred human intentions, planning systems that reason about a human's mental state rather than just their physical trajectory, and learning methods that let all of the above adapt on the fly rather than being hand-coded once and frozen.

We organize the discussion around five themes that recur across this literature, in roughly the order a system architect would encounter them: the cognitive-architecture tradition that supplies the representational backbone for robot reasoning; the perception and intention-recognition layer that turns sensor data into an estimate of what a human collaborator is doing; the trust, safety, and task-planning layer that decides what the robot should do given that estimate; the machine-learning and digital-twin methods that let the whole stack adapt to new tasks, new partners, and drifting conditions; and, finally, the open problems that the literature itself flags as unresolved. As with any synthesis of an active, fast-moving field, we have tried to be precise about what is established versus what is merely promising: cognitive-architecture integration into physical robots has a fairly long, well-documented track record; intention recognition has moved rapidly from rule-based heuristics to deep-learning pipelines with genuinely high reported accuracies on narrow benchmark tasks; and trust-aware planning and reinforcement-learning-based task allocation remain, by the admission of their own

authors, closer to promising demonstrations than deployed standards. We try to keep those distinctions visible throughout rather than flattening the literature into a single triumphant narrative.

It is also worth being explicit about scope. "Cognitive computing" is sometimes used loosely enough to mean almost any application of machine learning, and we do not use it that way here: we use it, consistent with its origin in the cognitive-architecture and symbolic-AI literature, to mean systems that maintain and reason over structured internal representations of a task and a partner's mental state — goals, intentions, beliefs about what the other agent knows — rather than systems that map sensor input directly to actuator output with no such intermediate representation. A robot that only maps input to output has no principled way to explain, predict, or negotiate around a human partner's behavior; a robot with even a simplified cognitive layer does, and the literature reviewed below is largely about how to build that layer cheaply and reliably enough to run on real collaborative hardware rather than only in simulation.

2. Cognitive Architecture Foundations for Robot Reasoning

Cognitive architectures predate collaborative robotics by decades, developed originally as computational theories of human cognition rather than as robot control systems. Soar, one of the two most influential such architectures, represents an agent's knowledge as production rules operating over a working memory, with a fixed decision cycle that selects and applies operators to move the agent toward its goals; Laird's introduction to the architecture lays out this rule-based, goal-driven cycle and traces how it has been extended over four decades to include episodic memory, reinforcement learning, and mechanisms for handling impasses when no rule directly applies [5]. ACT-R, the other dominant architecture, takes a related but distinct approach, splitting long-term knowledge into declarative memory (facts) and procedural memory (production rules), with activation-based retrieval that gives the architecture quantitative, often surprisingly accurate, predictions about human reaction times and error rates on cognitive tasks. Both architectures were built to model human cognition in the abstract; what makes them relevant to this article is that both have been explicitly embodied in physical robots to give those robots a comparable

reasoning substrate.

Trafton and colleagues' ACT-R/E is the clearest example of this embodiment. Rather than treating the cognitive architecture as a disembodied reasoning engine bolted onto a robot's perception and motor systems as an afterthought, ACT-R/E extends ACT-R's declarative and procedural memory system with spatial and perceptual modules tied directly to a physical robot's sensors and actuators, so that the same architecture that models a human's working memory limitations and retrieval dynamics also drives the robot's real-time perception-action loop [6]. The practical payoff the authors report is a robot that can perform tasks requiring perspective-taking — reasoning about what a human partner can see or reach from their physical vantage point, not just what the robot itself can see — because the architecture's spatial reasoning module is general enough to be queried from either agent's frame of reference. This is a small but genuinely important capability for collaboration: a robot that only models its own workspace has no principled way to hand an object to a person standing in an awkward position or to avoid blocking a view the person needs, and a purely reactive control system has no representation over which "what can the human see from there" is even a well-formed question.

The cost of this representational richness is computational and, at times, architectural rigidity. Both Soar and ACT-R impose a fairly specific cycle structure and knowledge representation, which is a feature when the goal is cognitive fidelity to human reasoning but can be a constraint when the goal is fast, robust behavior on constrained robot hardware operating under hard real-time deadlines. This tension — between the representational power a full cognitive architecture provides and the computational overhead it imposes — recurs throughout the more recent literature reviewed below, most of which adopts lighter-weight, more narrowly scoped cognitive layers (intention classifiers, task-specific planners with an explicit human model) rather than a full symbolic cognitive architecture, precisely because the narrower layer is easier to make real-time and easier to retrain for a new task. Freire and colleagues' socially adaptive cognitive architecture for industrial human-robot collaboration is a useful recent data point on this trade-off: rather than porting a general-purpose architecture like Soar or ACT-R

wholesale, the authors build a purpose-specific cognitive control loop, grounded in a neuroscience-inspired model of reinforcement and habit learning, that adapts a robot's collaborative behavior to an individual human partner's working style over repeated interactions in an industrial assembly setting [7]. Their framework is narrower in scope than a general cognitive architecture but is explicitly designed to run in real time on physical collaborative hardware, which is precisely the constraint that motivates most of the narrower architectures discussed in the sections that follow.

From Symbolic Reasoning to Learned Representations

A second thread within the cognitive-architecture literature, less concerned with faithfully modeling human cognition and more concerned with giving a robot a workable internal model of the collaborative task itself, centers on shared mental models. The concept, borrowed from human team-science research, holds that effective teams — human or mixed human-robot — perform better when members maintain compatible internal representations of the task, the roles, and each other's capabilities, rather than each member reasoning from an isolated, private model of the situation. Tabrez, Luebbbers, and Hayes's survey of mental modeling techniques in human-robot teaming organizes the computational approaches to this idea into a small number of families: techniques that explicitly represent and reason about what the human knows or believes (theory-of-mind-style approaches), techniques that infer a human's goals or preferences from observed behavior and fold that inference into the robot's own planning, and techniques that focus on making the robot's own internal state legible to the human, on the theory that a shared mental model requires legibility in both directions, not just robot-to-human inference [8]. Their survey is candid that most published systems address only one of these directions — human-to-robot inference — and that robot-to-human legibility, communicating the robot's plan or uncertainty in a way a human can quickly interpret, is comparatively underdeveloped, a gap we return to in the open-challenges section.

Vinanzi, Cangelosi, and Goerick's work on intention reading and trust operationalizes a related idea at a more concrete algorithmic level, combining a

probabilistic model of human intention with an explicit, updating trust estimate, so that the robot's confidence in its own intention inference is itself tracked and used to modulate how the robot acts on that inference [9]. This coupling of intention inference with trust is a recurring architectural pattern in the literature discussed below: a cognitive layer that only estimates what the human wants, without also estimating how confident that estimate is, tends to produce brittle behavior, because a wrong-but-confident inference and a wrong-but-uncertain inference should lead a well-designed robot to very different actions (the former to committing to an incorrect plan, the latter to seeking clarification or hedging). Whether that coupling is represented symbolically, as in the classical cognitive-architecture tradition, or learned end-to-end from data, as in most of the perception systems discussed next, the underlying requirement — a robot that tracks not just what it believes about the human but how much to trust that belief — is close to a structural invariant across this literature.

3. Perception and Intention Recognition

Whatever cognitive layer sits on top of a collaborative robot, it is only as good as the estimate of human state it receives from below, and the perception and intention-recognition literature has moved rapidly over the past several years from hand-engineered heuristics toward deep-learning pipelines that fuse multiple sensing modalities. Kekana and colleagues' review of human intention recognition frameworks in industrial collaborative robotics is, to our knowledge, the most comprehensive recent synthesis of this specific sub-literature, and it is useful precisely because it reports concrete, modality-specific accuracy figures rather than only qualitative claims about what deep learning "enables" [10]. Vision-based systems using hand or gesture tracking report accuracies approaching the ceiling on constrained recognition tasks — finger-joint-angle-based hand gesture recognition with a Leap Motion sensor reaching essentially perfect accuracy on its benchmark set, for instance — while systems that fuse gaze tracking with skeletal pose data for intention prediction report similarly high accuracy on their respective evaluation tasks. Force- and torque-based approaches, which infer intention from the physical interaction forces at a shared object or tool rather than from vision, report F1 scores in the high nineties with sub-second inference latency, a

combination that matters a great deal for real-time collaboration, where a technically accurate intention estimate that arrives too late to act on is not actually useful. Wearable inertial sensing, used for tasks like predicting when a human is about to reach for a handover, reports comparably high prediction accuracy with low variance across trials, and deep sequence models — LSTMs and related recurrent architectures applied to hand-motion trajectories — report somewhat lower but still substantial validation accuracy on motion-prediction tasks, reflecting the intrinsically harder problem of predicting a continuous trajectory rather than classifying a discrete gesture. Multimodal fusion approaches, combining several of these signal types through a probabilistic model, report accuracy and precision figures in the high eighties to low nineties, generally lower than the best single-modality result on its own narrow benchmark but, the review's authors argue, more robust across the range of conditions a real industrial floor actually presents, where any single sensing modality is liable to fail or degrade (occluded cameras, noisy force signals, inertial drift) under conditions a laboratory benchmark does not fully capture.

Table 1: Reported accuracy figures for representative sensing modalities used for human intention recognition in industrial collaborative robotics, as synthesised in Kekana et al.'s review [10].

Modality / Technique	Signal Basis	Reported Metric	Reported Value
Vision, finger-joint hand gesture (Leap Motion)	Optical hand tracking	Classification accuracy	100% on benchmark gesture set
Vision, gaze + skeletal pose fusion	Gaze tracking with body pose	Intention prediction accuracy	100% on evaluated task
Force/torque sensing	Physical interaction force at shared object	F1 score / inference time	98.14% F1, 0.85 s latency
Wearable IMU, handover prediction	Fuzzy rule-based motion sensing	Prediction accuracy	99.6% (SD 0.008)
Deep learning, LSTM hand-motion prediction	Sequential trajectory modelling	Validation accuracy	89%
Multimodal Bayesian fusion	Combined vision + context cues	Accuracy / precision	89.55% / 91.80%

Mavsar, Simonič, and Ude's more recent work on human intention recognition with deep LSTM and transformer networks pushes this line of work toward architectures explicitly designed for the sequential,

temporally extended nature of collaborative tasks, where a human's intention is not a single instantaneous gesture but an unfolding sequence of actions whose meaning often only becomes disambiguated partway through [11]. Transformer-based sequence models, with their attention mechanism explicitly designed to weigh relevant context across a variable-length history, are a natural fit for this problem, and the authors report that their approach supports real-time intention recognition suitable for closed-loop human-robot collaboration rather than only offline classification of pre-segmented action clips — an important distinction, because a large share of the earlier intention-recognition literature evaluated models on cleanly segmented benchmark clips rather than on the continuous, unsegmented sensor stream a deployed system actually has to process.

It is worth pausing on why so many of these modalities are pursued in parallel rather than the field simply converging on whichever single sensor type reports the highest number. Vision is cheap, non-contact, and captures a wide field of view, but is vulnerable to occlusion whenever a tool, a workpiece, or the robot's own arm passes between the camera and the human, which happens often in a genuinely shared workspace rather than a laboratory bench. Force and torque sensing is comparatively occlusion-proof and fast, since it measures the physical interaction directly rather than inferring it from an image, but it only becomes informative once contact has already begun, which is too late for some anticipatory behaviors a robot might otherwise want to take. Wearable inertial sensing sidesteps both of these limitations by moving the sensor onto the human's body, at the cost of requiring the human to wear something, which is a nontrivial adoption barrier in many industrial settings regardless of how accurate the underlying signal turns out to be. Multimodal fusion is the literature's answer to this set of complementary weaknesses rather than an attempt to beat any single modality's benchmark number outright, and it is worth being direct about what these accuracy figures do and do not establish. Nearly all of them come from constrained laboratory or benchmark evaluations with a fixed, relatively small set of gesture or action classes, controlled lighting and camera placement, and a limited pool of human subjects; none of them, as far as the reviewed sources report, come from long-duration deployment on a factory floor with the full diversity of worker heights, clothing, working styles, and incidental

motion that a real industrial environment presents. This is precisely the gap the perception literature itself is most candid about, and it is a gap that recurs, in a different form, in essentially every layer of the cognitive-computing stack discussed in this article: benchmark performance is consistently strong, deployment-scale validation is comparatively thin.

4. Trust, Safety, and Human-Aware Task Planning

An accurate intention estimate is only useful if the robot's planning and control layer actually does something sensible with it, and a second major strand of this literature addresses exactly that: how should a robot plan its actions — what to do, when to do it, how much autonomy to claim in a given moment — given an estimate, necessarily imperfect, of what a human partner is doing and intends? Safety is the most foundational layer of this question, and industrial standards now formalize it explicitly rather than leaving it to individual system designers. Villani and colleagues' survey catalogs the collaborative operation modes defined by international safety standards for industrial robots — safety-rated monitored stop, hand-guiding, speed and separation monitoring, and power and force limiting — and argues that the choice among these modes is not merely a safety-engineering decision but a design decision that constrains what kind of cognitive collaboration is even possible in a given deployment, because a system relying purely on hard-coded speed and separation limits has comparatively little use for a rich human-intention model, while a system that wants to actively hand objects to, or receive objects from, a human partner in close proximity needs exactly the kind of intention and trust modeling discussed above to do so safely [3].

Trust itself has been studied as a quantitative, modelable variable in its own right, largely independent of any particular safety standard. Hancock and colleagues' meta-analysis of the factors affecting trust in human-robot interaction, synthesizing correlational and experimental studies published up to that point, reports an overall correlational effect size of mean $r = 0.26$ and an overall experimental effect size of mean $d = 0.71$ for factors influencing trust, and — more usefully for system design — finds that robot performance-related factors (reliability, predictability, dependability) and robot attributes were considerably stronger predictors of trust

than either environmental factors or factors relating to the human operator themselves [12]. That asymmetry has a direct architectural implication that recurs across the more recent literature: if trust is driven primarily by the robot's observable performance rather than by static attributes of the environment or the human, then a cognitive-computing framework aiming to sustain trust over time needs mechanisms for the robot to behave predictably and communicate its intentions clearly, rather than mechanisms aimed primarily at, say, tailoring its appearance or persona to the individual human partner. Khavas's survey of trust-modeling approaches in human-robot interaction extends this line of work by cataloging the computational mechanisms — Bayesian trust estimation, reinforcement-learning-based trust updates, and hybrid performance-and-disposition models — that have been proposed to track and predict a human's trust in a robot as it evolves over the course of an interaction, rather than treating trust as a fixed, one-time-measured quantity [13].

The underlying intuition that anticipation, not just correctness, drives good collaboration goes back further than most of the trust-modeling work above. Hoffman and Breazeal's early study of anticipatory action in human-robot teamwork found that a robot that acts proactively on a predicted next step, rather than waiting to react only after the human's action is complete, measurably improves objective task efficiency and the human partner's subjective sense of team fluency, even when the anticipatory action carries some risk of being wrong [14]. That finding anchors much of what the more recent intention-recognition and trust-modeling literature is, in effect, trying to make robust and general: anticipation only pays off if the underlying prediction is good enough, and often enough, to outweigh the cost of the occasional wrong guess, which is precisely the calculus a robot's coupled intention-and-trust estimate needs to get right. Chen and colleagues' trust-aware decision-making framework operationalizes this coupling directly inside the robot's action-selection process: rather than planning purely to complete the task efficiently, the robot's planner explicitly models the human's evolving trust in it as part of the state it is optimizing over, learning both a model of how the human's trust responds to the robot's actions and a policy that accounts for the fact that an action which is efficient in isolation may erode trust enough to degrade

the collaboration's longer-run performance [15]. This is a meaningfully different design philosophy from treating trust as an outcome to be measured after the fact in a user study; it treats trust as a state variable the planner reasons about online, alongside more conventional state variables like task progress and human position. The POMDP (partially observable Markov decision process) formalism has proven a natural mathematical framework for this kind of reasoning, precisely because it was built to handle decision-making under exactly the kind of hidden-state uncertainty that a robot's estimate of human intention or human trust represents. Zhang and Lin's POMDP-based supervisory control framework for human-robot collaboration provides formal performance guarantees for a robot operating under this kind of uncertainty, an important complement to the more empirically driven trust and intention literature, because it addresses the question of what can be proven about a collaborative system's behavior rather than only what has been observed to work well in a specific experimental setting [16].

Communication-Free and Planning-Level Adaptation

A related but distinct planning problem is how a robot should adapt its own task allocation and sequencing to a specific human partner's working style, without relying on explicit verbal communication or negotiation, since requiring a human to constantly instruct the robot defeats much of the point of collaboration in the first place. Ramachandruni, Kent, and Chernova's UHTP framework addresses this directly with a user-aware hierarchical task planner that infers a human collaborator's preferences and capabilities from observed behavior during the task itself, and adapts the robot's own subtask selection and sequencing to complement that inferred profile, all without requiring the human to communicate preferences explicitly [17]. The hierarchical structure matters here for largely the same reason it matters in classical AI planning more broadly: it lets the system reason about coarse-grained role allocation (who handles which subtask) separately from fine-grained motion-level adaptation, which keeps the online inference problem tractable even though the full space of possible human working styles and task decompositions is combinatorially large. Read together with the trust-aware and POMDP-based planning work above, this body of research supports a fairly specific claim: task planning for human-robot collaboration is no longer treated in the recent literature as a pure

scheduling or motion-planning problem to be solved once the human's intentions are known, but as an online inference-and-adaptation problem in which the robot's model of the human — their intentions, their trust, their working preferences — is itself a first-class part of the planning state, continuously updated rather than assumed fixed.

5. Machine Learning, Reinforcement Learning, and Digital Twins for Adaptive Collaboration

Beneath the intention-recognition and trust-aware planning layers described above sits a broader machine-learning literature concerned with how a collaborative robot should actually learn, and keep learning, the behaviors those layers depend on. Semeraro, Griffiths, and Cangelosi's systematic review of machine learning applied to human-robot collaboration is a useful empirical anchor for how this sub-field has actually distributed its effort: across the 45 papers the review selected for detailed analysis, the authors identify 23 cases applying supervised learning techniques, 18 cases applying reinforcement learning, and 12 cases applying unsupervised learning, with several papers combining more than one technique [18]. That distribution is informative in its own right — supervised learning, generally applied to perception and intention-classification problems of the kind discussed in the previous section, remains the single most common technique, while reinforcement learning, generally applied to the sequential decision-making and task-allocation problems discussed just above, is a close second and has grown rapidly as a share of the field's output in more recent years covered by the review. The review's authors emphasize one methodological finding in particular that is worth foregrounding: because human-robot collaboration is inherently a sequential, time-extended process rather than a set of independent, identically distributed samples, machine learning methods that explicitly model time dependencies — recurrent architectures, sequence-to-sequence models, and reinforcement learning formulations with explicit state transitions over time — consistently outperform methods that treat each observation independently, a finding that echoes, from the ML methodology side, the same emphasis on temporally extended reasoning that motivated the transformer- and LSTM-based intention recognition work discussed earlier [18,11].

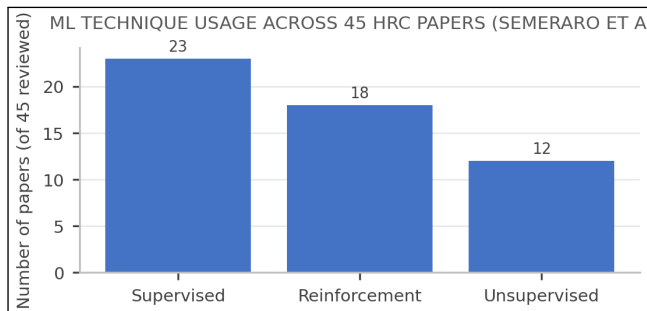


Figure 1: Distribution of machine learning technique types across the 45 human-robot collaboration papers analysed in Semeraro et al.'s systematic review -- supervised learning (mostly perception and intention classification) remains the most common approach, with reinforcement learning (mostly task allocation and adaptive control) a close second [18].

Reinforcement learning applied specifically to task allocation and scheduling in assembly settings has developed into a reasonably mature sub-literature in its own right. Zhang and colleagues propose a reinforcement-learning method for dynamically allocating subtasks between a human and a robot during an assembly process, allowing the division of labor to adapt online as the task unfolds rather than being fixed at design time by a human engineer, and report that the approach shows clear potential for improving the flexibility of task division compared to static, pre-assigned allocation schemes [19]. Akkaladevi and colleagues' earlier interactive reinforcement-learning framework for collaborative assembly processes takes a complementary approach, structuring the learning problem so that the robot can incorporate corrective human feedback directly into its policy updates during the task itself, rather than requiring a separate, offline training phase before deployment — a design choice that matters practically because collaborative assembly tasks are frequently too varied, and too costly to interrupt for retraining, to tolerate a long offline learning phase before the robot becomes useful [20]. Read alongside each other, these two lines of reinforcement-learning work sit at somewhat different points on a spectrum: one treats the robot's policy as something to be adapted continuously online in response to the evolving task state, the other as something to be adapted continuously online in response to explicit human correction, and a fully mature deployed system arguably needs both kinds of adaptivity operating together. A related and considerably older learning paradigm, learning from demonstration, addresses a slightly different problem: rather than learning a policy from scratch through trial-and-error interaction, the robot learns directly from

examples of the task performed by a human. Argall, Chernova, Veloso, and Browning's foundational survey of robot learning from demonstration lays out the core design choices this paradigm still turns on — how demonstrations are recorded, how the mapping from a human demonstrator's actions to a robot's own actuators is handled when the two bodies differ, and how a policy learned from a necessarily limited set of demonstrated examples is made to generalize to states the demonstrator never showed it [21]. For human-robot collaboration specifically, learning from demonstration and reinforcement learning are increasingly used together rather than as competitors: a policy initialized from a handful of human demonstrations, then refined through continued reinforcement-learning-style interaction once deployed, converges faster and behaves more predictably during the early, still-untrained phase of collaboration than a policy trained from reinforcement signals alone.

Digital twins represent a more recent and, in our reading, still less settled addition to this stack, but one that several groups have identified as a natural complement to the learning-based methods above precisely because reinforcement learning and other data-hungry methods benefit enormously from a safe, fast simulation environment in which to train before any policy is deployed on physical hardware around a real person. Ramasubramanian and colleagues' review of digital twins for human-robot collaboration in manufacturing frames the digital twin not merely as a passive visualization or monitoring tool but as an active component of the collaboration architecture itself: a continuously synchronized virtual replica of the physical cell that can be used to test candidate robot behaviors, including learned policies, in simulation before they are executed near a human, and to maintain an updating estimate of both the robot's and the human's state for the planning and trust-modeling layers discussed earlier to consume [22]. Their review is candid that most published digital-twin systems for human-robot collaboration remain at a proof-of-concept or laboratory-demonstration stage rather than full production deployment, and that synchronizing a digital twin's model of a human worker — as opposed to the comparatively well-understood problem of synchronizing a model of the robot itself — with the physical human's actual state in real time remains a substantially harder and less mature engineering

problem, echoing the same human-modeling difficulty that runs through the intention-recognition and trust literature discussed earlier in this article. Growing industrial demand for exactly this kind of system is itself part of the evidentiary picture here: Keshvarparast and colleagues' large-scale literature review of collaborative robots in manufacturing and assembly, screening 201 papers across two major databases, notes that the global collaborative robot market's share of the broader industrial robotics market grew from roughly 5% in 2018 to a projected 13% by 2022, alongside market-size estimates rising from around 700 million US dollars in 2021 toward a projected nearly two billion US dollars by 2030 — figures that, whatever one makes of market forecasting generally, at minimum indicate that the deployment context motivating this entire literature is not a niche or shrinking one [23]. Weidemann and colleagues' companion literature review of collaborative robotics trends under the "Work 4.0" framing reaches a broadly consistent conclusion from an independent survey of the field, situating this growth within a wider set of drivers -- an aging manufacturing workforce, demand for smaller and more flexible production batches, and a policy push in several economies toward keeping skilled human labor in the loop rather than replacing it outright -- that helps explain why cognitive-computing investment in this specific application has accelerated rather than plateaued in recent years [24].

6. Open Challenges

Several gaps recur across the literature synthesized here clearly enough to be worth stating plainly rather than leaving implicit. First, the evaluation-realism gap already flagged in the perception section applies almost across the board: cognitive architectures like ACT-R/E have a genuine, multi-decade track record of embodiment on physical robots, but most of the intention-recognition, trust-modeling, and reinforcement-learning results surveyed here are reported from constrained laboratory studies, with a limited number of human subjects, a fixed and relatively small task vocabulary, and evaluation timescales measured in single sessions rather than the weeks or months of continued use that would reveal whether a learned trust model or an adapted task allocation policy remains useful, or degrades, as a human partner's own behavior and expectations shift over sustained use [10,8]. Second, and closely related, the legibility half of

the shared-mental-model problem — communicating the robot's own state, plan, or uncertainty back to the human in a way the human can quickly and correctly interpret — remains considerably less developed than the human-to-robot inference half; Tabrez, Luebbers, and Hayes make this asymmetry explicit in their survey, and it has not, to our knowledge, closed substantially in the more recent literature reviewed here [8].

Third, the relationship between the classical, symbolic cognitive-architecture tradition and the more recent, largely deep-learning-based intention-recognition and reinforcement-learning literature is, candidly, under-integrated. Architectures like Soar and ACT-R offer interpretable, inspectable reasoning traces attractive from a safety and certification standpoint, the same standpoint that motivates the ISO-standard collaborative operation modes discussed above, but comparatively few of the high-accuracy deep-learning intention-recognition systems reviewed here are built to interoperate with, or be audited by, that kind of symbolic reasoning layer [5,3]. Freire and colleagues' socially adaptive architecture is a partial exception, but it remains closer to the leading edge of this integration than to an established pattern [7]. Fourth, trust modeling and safety-standard compliance have developed largely as separate literatures with different methodological traditions, and a framework that jointly optimizes for both, rather than treating safety as a hard constraint and trust as a separately optimized soft objective, remains more aspiration than accomplished result in the sources reviewed here [12,16].

Fifth, and more practically, digital twins and reinforcement-learning-based task allocation both depend on models of human behavior that are considerably harder to build and keep synchronized with reality than the corresponding models of the robot itself, and this asymmetry is, in our reading, the single most consistent theme across every layer of this literature: predicting the robot is a comparatively solved problem, predicting the human collaborating with it is not [22,19]. None of these gaps undermine the case that cognitive computing frameworks are a productive, increasingly well-evidenced direction for adaptive human-robot collaboration — independent groups have converged on broadly similar architectural patterns — but future claims of deployment readiness should be read against these gaps, not only against the more

polished benchmark numbers.

7. Conclusion

The literature surveyed here supports a reasonably specific claim about where cognitive computing for human-robot collaboration currently stands. The field has a genuine architectural lineage running from general-purpose cognitive architectures like Soar and ACT-R, embodied directly in physical robots through extensions like ACT-R/E, through increasingly narrow and computationally tractable cognitive layers purpose-built for specific collaborative tasks [5,6,7]. It has a perception and intention-recognition sub-literature that has moved from rule-based heuristics to deep sequence models with genuinely high reported accuracies on constrained benchmarks, even as the field itself acknowledges that those benchmarks fall well short of full deployment realism [10,11]. It has a planning and decision-making layer that increasingly treats trust, human intention, and human working preference not as fixed inputs but as evolving hidden state to be estimated and reasoned about online, with formal frameworks like POMDPs providing at least partial guarantees about behavior under that uncertainty [15,16,17]. And it has a rapidly growing machine-learning and digital-twin literature supplying the adaptive backbone that lets all of the above continue learning after deployment rather than being frozen at design time [18,19,22].

What ties these threads together, more than any single shared algorithm, is a shared commitment to giving the robot an explicit, updating internal model of its human partner — their intentions, their trust, their working style — rather than treating the human as simply another obstacle in the workspace or another set of joint-torque constraints on the robot's own motion. That commitment is what distinguishes a cognitive-computing framework for human-robot collaboration from a conventional collaborative-robot control stack, and the evidence surveyed here, spanning decades-old symbolic architectures and this year's transformer-based intention recognizers alike, suggests it is a productive one. Making it fully deployment-ready, rather than merely well-evidenced in the laboratory, depends on closing the gaps named in the previous section: evaluating over realistic deployment timescales and worker populations rather than single laboratory sessions, building genuine two-way legibility rather than one-directional inference, and integrating the

interpretability and safety guarantees of the symbolic cognitive-architecture tradition with the empirical performance of modern learned perception and planning. None of these are reasons to discount the pattern; they are, taken together, closer to the field's own honest roadmap for what comes next, and the growing industrial demand for collaborative robots documented in the recent literature suggests that roadmap will keep getting tested against real deployments rather than only against further laboratory benchmarks [23].

8. References

1. Gutierrez-Garcia, J.O., Lopez-Neri, E. (2015) Cognitive Computing: A Brief Survey and Open Research Challenges. Proceedings of the 3rd International Conference on Applied Computing and Information Technology / 2nd International Conference on Computational Science and Intelligence (ACIT-CSI). DOI: [10.1109/ACIT-CSI.2015.64](https://doi.org/10.1109/ACIT-CSI.2015.64)
2. Goodrich, M.A., Schultz, A.C. (2008) Human-Robot Interaction: A Survey. Foundations and Trends in Human-Computer Interaction, 1(3): 203-275. DOI: [10.1561/11000000005](https://doi.org/10.1561/11000000005)
3. Villani, V., Pini, F., Leali, F., Secchi, C. (2018) Survey on Human-Robot Collaboration in Industrial Settings: Safety, Intuitive Interfaces and Applications. Mechatronics, 55: 248-266. DOI: [10.1016/j.mechatronics.2018.02.009](https://doi.org/10.1016/j.mechatronics.2018.02.009)
4. Singh, S., Rantanen, E., Heard, J. (2026) Human-Robot Teaming: A Comprehensive Survey on Collaboration, Communication, and Cognition. ACM Transactions on Human-Robot Interaction, 15(2): 1-48. DOI: [10.1145/3776548](https://doi.org/10.1145/3776548)
5. Laird, J.E. (2022) Introduction to the Soar Cognitive Architecture. arXiv preprint. DOI: [10.48550/arXiv.2205.03854](https://doi.org/10.48550/arXiv.2205.03854)
6. Trafton, J.G., Hiatt, L.M., Harrison, A.M., Tamborello, F.P., Khemlani, S.S., Schultz, A.C. (2013) ACT-R/E: An Embodied Cognitive Architecture for Human-Robot Interaction. Journal of Human-Robot Interaction, 2(1): 30-55. DOI: [10.5898/JHRI.2.1.Trafton](https://doi.org/10.5898/JHRI.2.1.Trafton)
7. Freire, I.T., Guerrero-Rosado, O., Amil, A.F., Verschure, P.F.M.J. (2024) Socially Adaptive Cognitive Architecture for Human-Robot Collaboration in Industrial Settings. Frontiers in Robotics and AI, 11: 1248646. DOI: [10.3389/frobt.2024.1248646](https://doi.org/10.3389/frobt.2024.1248646)
8. Tabrez, A., Luebbbers, M.B., Hayes, B. (2020) A Survey of Mental Modeling Techniques in Human-Robot Teaming. Current Robotics Reports, 1(4): 259-267. DOI: [10.1007/s43154-020-00019-0](https://doi.org/10.1007/s43154-020-00019-0)
9. Vinanzi, S., Cangelosi, A., Goerick, C. (2021) The Collaborative

- Mind: Intention Reading and Trust in Human-Robot Interaction. *iScience*, 24(2): 102130. DOI: [10.1016/j.isci.2021.102130](https://doi.org/10.1016/j.isci.2021.102130)
10. Kekana, M., Du, S., Steyn, N., Benali, A., Djerroud, H. (2025) A Review of Human Intention Recognition Frameworks in Industrial Collaborative Robotics. *Robotics*, 14(12): 174. DOI: [10.3390/robotics14120174](https://doi.org/10.3390/robotics14120174)
 11. Mavsar, M., Simonič, M., Ude, A. (2025) Human Intention Recognition by Deep LSTM and Transformer Networks for Real-Time Human-Robot Collaboration. *Frontiers in Robotics and AI*, 12: 1708987. DOI: [10.3389/frobt.2025.1708987](https://doi.org/10.3389/frobt.2025.1708987)
 12. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J., Parasuraman, R. (2011) A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, 53(5): 517-527. DOI: [10.1177/0018720811417254](https://doi.org/10.1177/0018720811417254)
 13. Khavas, Z.R. (2020) Modeling Trust in Human-Robot Interaction: A Survey. *arXiv preprint*. DOI: [10.48550/arXiv.2011.04796](https://doi.org/10.48550/arXiv.2011.04796)
 14. Hoffman, G., Breazeal, C. (2007) Effects of Anticipatory Action on Human-Robot Teamwork Efficiency, Fluency, and Perception of Team. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI '07)*: 1-8. DOI: [10.1145/1228716.1228718](https://doi.org/10.1145/1228716.1228718)
 15. Chen, M., Nikolaidis, S., Soh, H., Hsu, D., Srinivasa, S. (2020) Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning. *ACM Transactions on Human-Robot Interaction*, 9(2): 1-23. DOI: [10.1145/3359616](https://doi.org/10.1145/3359616)
 16. Zhang, X., Lin, H. (2019) Performance Guaranteed Human-Robot Collaboration with POMDP Supervisory Control. *Robotics and Computer-Integrated Manufacturing*, 57: 59-72. DOI: [10.1016/j.rcim.2018.10.011](https://doi.org/10.1016/j.rcim.2018.10.011)
 17. Ramachandrani, K., Kent, C., Chernova, S. (2024) UHTP: A User-Aware Hierarchical Task Planning Framework for Communication-Free, Mutually-Adaptive Human-Robot Collaboration. *ACM Transactions on Human-Robot Interaction*, 13(3): 1-27. DOI: [10.1145/3623387](https://doi.org/10.1145/3623387)
 18. Semeraro, F., Griffiths, A., Cangelosi, A. (2023) Human-Robot Collaboration and Machine Learning: A Systematic Review of Recent Research. *Robotics and Computer-Integrated Manufacturing*, 79: 102432. DOI: [10.1016/j.rcim.2022.102432](https://doi.org/10.1016/j.rcim.2022.102432)
 19. Zhang, R., Lv, Q., Li, J., Bao, J., Liu, T., Liu, S. (2022) A Reinforcement Learning Method for Human-Robot Collaboration in Assembly Tasks. *Robotics and Computer-Integrated Manufacturing*, 73: 102227. DOI: [10.1016/j.rcim.2021.102227](https://doi.org/10.1016/j.rcim.2021.102227)
 20. Akkaladevi, S.C., Plasch, M., Maddukuri, S., Eitzinger, C., Pichler, A., Rinner, B. (2018) Toward an Interactive Reinforcement Based Learning Framework for Human Robot Collaborative Assembly Processes. *Frontiers in Robotics and AI*, 5: 126. DOI: [10.3389/frobt.2018.00126](https://doi.org/10.3389/frobt.2018.00126)
 21. Argall, B.D., Chernova, S., Veloso, M., Browning, B. (2009) A Survey of Robot Learning from Demonstration. *Robotics and Autonomous Systems*, 57(5): 469-483. DOI: [10.1016/j.robot.2008.10.024](https://doi.org/10.1016/j.robot.2008.10.024)
 22. Ramasubramanian, A.K., Mathew, R., Kelly, M., Hargaden, V., Papakostas, N. (2022) Digital Twin for Human-Robot Collaboration in Manufacturing: Review and Outlook. *Applied Sciences*, 12(10): 4811. DOI: [10.3390/app12104811](https://doi.org/10.3390/app12104811)

23. Keshvarparast, A., Battini, D., Battaia, O., Pirayesh, A. (2024) Collaborative Robots in Manufacturing and Assembly Systems: Literature Review and Future Research Agenda. *Journal of Intelligent Manufacturing*, 35(5): 2065-2118. DOI: [10.1007/s10845-023-02137-w](https://doi.org/10.1007/s10845-023-02137-w)
24. Weidemann, C., Mandischer, N., van Kerkom, F., Corves, B., Hüsing, M., Kraus, T., Garus, C. (2023) Literature Review on Recent Trends and Perspectives of Collaborative Robotics in Work 4.0. *Robotics*, 12(3): 84. DOI: [10.3390/robotics12030084](https://doi.org/10.3390/robotics12030084)

9. Declarations

Funding

This literature synthesis received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

No new data were generated for this article, as it is a literature synthesis. All studies discussed are cited in the reference list, and readers should consult the original publications for their underlying data and code availability statements.

Ethical Approval

Not applicable. This article is a literature synthesis and did not involve human participants, animal subjects, or primary data collection.