Review | Open Access

Intelligence & Robotics

Yu. *Intell. Robot.* **2026**, *6*(3), 544-59 **DOI:**10.20517/ir.2026.26

A survey on deep reinforcement learning for human-robot interaction

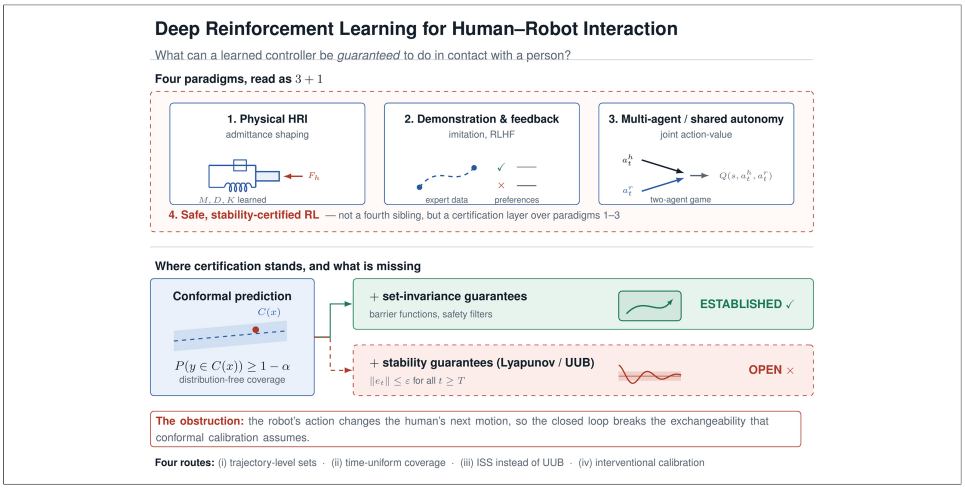
Wen Yu

Keywords:
Deep reinforcement learning, human-robot interaction, safe reinforcement learning, admittance control, conformal prediction

Citation: Yu, W. A survey on deep reinforcement learning for human-robot interaction. *Intell. Robot.* **2026**, *6*(3), 544-59. https://dx.doi.org/10.20517/ir.2026.26

Received: 13 Jul 2026
First Decision: 4 Aug 2026
Revised: 6 Aug 2026
Accepted: 17 Aug 2026
Published: 28 Aug 2026

Academic Editor:
Hao Zhang
Copy Editor:
Pei-Yun Wang
Production Editor:
Pei-Yun Wang



Abstract

Deep reinforcement learning (DRL) is reshaping human-robot interaction (HRI) control beyond fixed admittance/impedance laws. This survey bridges classical HRI control and DRL, then organises recent work into four paradigms: physical HRI via admittance shaping, learning from demonstration and feedback, multi-agent and shared-autonomy control, and safe, stability-certified reinforcement learning (RL). We further discuss emerging directions, language-conditioned policies, digital-twin-based sim-to-real transfer, and residual RL, and argue that, while conformal-prediction bounds are now routinely combined with set-invariance guarantees such as barrier functions in HRI, their pairing with Lyapunov/uniform ultimate boundedness (UUB) stability guarantees remains obstructed by the loss of exchangeability in closed loop, which we identify as the field's least developed dimension. We close with open challenges in sample efficiency, non-stationary human behaviour, generalisation across users, and benchmarking.

1. INTRODUCTION

Human-robot interaction (HRI) control spans two coupled layers: a physical layer, in which a robot exchanges forces, positions, and motion with a human partner in direct or teleoperated contact, and a cognitive/social layer, in which the robot must

Departamento de Control Automatico, CINVESTAV-IPN (National Polytechnic Institute), Mexico City 07360, Mexico.

Correspondence to: Prof. Wen Yu, Departamento de Control Automatico, CINVESTAV-IPN (National Polytechnic Institute), Mexico City 07360, Mexico. E-mail: wen.yu@cinvestav.mx

interpret intent, adapt to preferences, and coordinate actions with a human whose behaviour is only partially predictable. Classical solutions to the physical layer, impedance and admittance control[1], render the robot's response to contact forces through a fixed or scheduled mechanical relationship. These schemes are provably stable, but they are built on a model of the environment (and implicitly, of the human) that is rarely accurate: human arm stiffness, reaction time, and intent all vary across users and across a single task[2].

Deep reinforcement learning (DRL) offers an alternative: rather than hand-specifying an interaction law, the robot learns a policy directly from interaction data, allowing the impedance/admittance behaviour itself to become adaptive and human-specific[3,4]. This shift from designed to learned interaction control is the central subject of this survey.

The shift is not free, and the price is paid in a currency that HRI can least afford. A learned interaction law is a black box placed in mechanical contact with a person, so the two questions that classical control answered by construction, will it remain stable and will it remain safe, must now be answered statistically, from finite data, about a policy that continues to change after deployment. Three consequences organise this survey. First, sample efficiency stops being an engineering convenience and becomes an ethical constraint, because every sample is a physical interaction with a human body; this is what makes demonstration data, offline pre-training, and simulation transfer central rather than optional. Second, the environment is non-stationary in a specific and awkward way: the human adapts to the robot while the robot adapts to the human, so the stationarity assumption underlying the Markov decision process is violated by the very success of the learning process. Third, guarantees must be stated in a form that survives learning, which is why uncertainty quantification and set-invariance methods have moved from the periphery of this literature to its centre in the last three years.

The same pattern, a perception-decision-action loop whose reliability must be certified because a person is on the receiving end, is visible well beyond manipulation. Reviews of embodied AI in clinical settings organise the field along exactly this axis and reach a similar conclusion about the translation bottleneck, namely that the barrier is evaluation and standardisation rather than raw capability[5]. We draw on that parallel where it is informative, while keeping the scope of this survey on physical and cognitive HRI control.

Four developments justify a new survey rather than a simple update of our 2021 book Human-Robot Interaction Control Using Reinforcement Learning[6]: (i) safe and stability-certified RL has matured from Lyapunov-style heuristics into formal H_2/uniform ultimate boundedness (UUB) guarantees[7]; (ii) conformal-prediction-based uncertainty quantification has entered control design[8]; (iii) large language and vision-language-action models now condition manipulation policies on natural-language instruction[9,10]; and (iv) residual and hybrid reinforcement learning (RL), learning a correction on top of a model-based controller rather than a policy from scratch, has become the dominant paradigm for sample-efficient, safety-preserving deployment[4,11].

Table 1 organises the literature underlying this survey into nine paradigms that trace the field's progression from classical, fixed-law HRI control to certified and language-conditioned DRL. This classification also fixes the vocabulary used throughout the sections that follow.

The remainder of this survey is organised as follows. Section 2 bridges classical HRI control and DRL. Section 3 proposes a four-part classification of DRL approaches for HRI. Section 4 discusses emerging directions, Section 5 lists open challenges, and Section 6 concludes.

Table 1. From classical HRI control to deep RL: a classification of the literature underlying this survey

Paradigm	Ref.	Key contribution
1. Classical impedance/admittance control	[1]	Fixed mechanical interaction laws [Equation (1)]; foundation of physical HRI, provably stable but not adaptive
2. Human-behaviour modeling and admittance learning	[3,6,12-15]	Parametrised, human-in-the-loop admittance/teleoperation models; book-length synthesis of pre-2021 state of the art
3. Hybrid position/force RL under contact uncertainty	[4,11]	Blends model-based force control with RL compensation when the environment/contact model is unknown
4. Learning from demonstration and human feedback	[16-18]	Reduces sample complexity and improves early-training safety using expert trajectories or human preference signals
5. Multi-agent/shared-autonomy RL	[19-22]	Treats robot and human as coupled agents in redundant/task-space control, rather than robot-only optimisation
6. Safe and stability-certified RL	[7,8,23-30]	Formal H_2/UUB/barrier-function guarantees and conformal-prediction uncertainty bounds for learned HRI policies
7. General DRL-for-robotics foundations	[31,32]	Establish core DRL-robotics methodology (policy search, end-to-end visuomotor control) HRI-specific work builds on
8. Contemporary DRL-HRI surveys (2023-2025)	[33-35]	Map the current DRL-robotics landscape; HRI treated as one sub-topic among several, motivating a dedicated survey
9. Emerging: LLM/VLA-conditioned and residual RL	[9,10]	Instruction-following manipulation policies and residual/hybrid discrete-continuous RL for sample-efficient deployment

HRI: Human-robot interaction; RL: reinforcement learning; UUB: uniform ultimate boundedness; DRL: deep reinforcement learning; LLM: large language model; VLA: vision-language-action.

2. FROM CLASSICAL HRI CONTROL TO DEEP RL

Admittance and impedance control regulate the dynamic relationship between interaction force F_h (applied by the human) and robot motion. A standard admittance model relates force to a commanded position correction x_d through a virtual mass-damper-spring system,

$$M\ddot{x}_d + D\dot{x}_d + Kx_d = F_h, \tag{1}$$

where M, D, and K are the designer-chosen virtual inertia, damping, and stiffness[1]. Stability of Equation (1) is straightforward when F_h is bounded and slowly varying, but M, D, K are typically fixed offline; they do not adapt to a specific human's stiffness, fatigue, or intent. A large body of work addresses this by scheduling the gains online: variable admittance schemes modulate D from measured interaction power or estimated user intent[12,13], and variable impedance learning treats the stiffness profile itself as the object to be adapted[3,14]. In all of these the interaction law still relies on a parametric structure fixed at design time.

Reinforcement learning removes that structural assumption. The HRI problem is cast as a Markov decision process $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where the state $s_t \in \mathcal{S}$ includes robot and (where observable) human kinematic/force data, the action $a_t \in \mathcal{A}$ is a motion or force command, and the policy $\pi_\theta(a_t \mid s_t)$ is trained to maximize expected discounted return,

$$\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right], \tag{2}$$

with the reward $r(s_t, a_t)$ typically shaped to penalize excessive interaction force, task error, and unsafe configurations. Three properties of the human partner make Equation (2) attractive over the fixed model in Equation (1):

• **Unknown human dynamics.** The mapping from robot behaviour to human response is rarely known in closed form; RL treats it as part of the environment transition P rather than requiring an explicit model. Evidence from human motor control supports this treatment: people co-adapt force and impedance rather

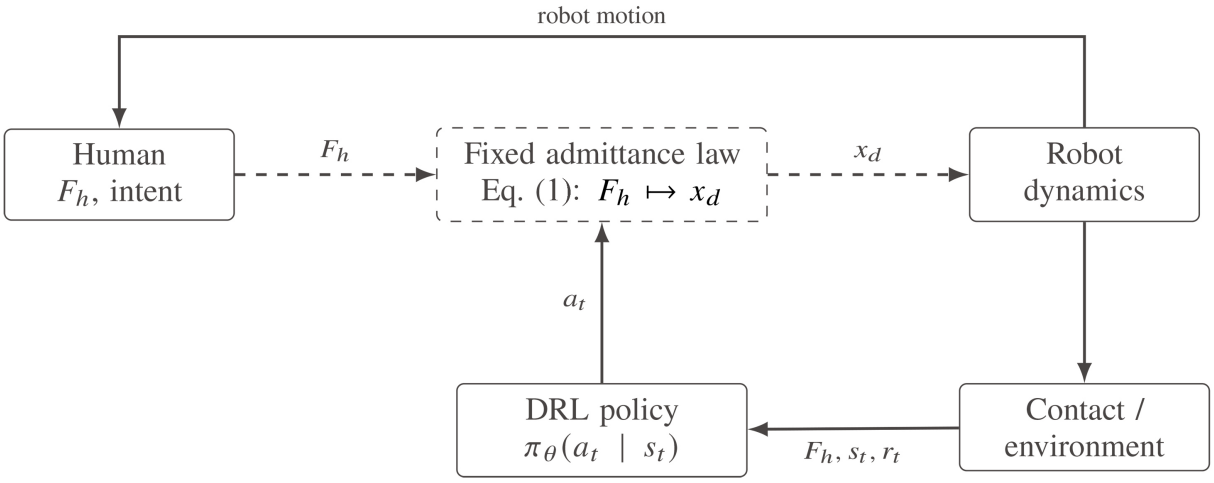


Figure 1. Classical admittance control (dashed) *vs.* DRL-based HRI control (solid). In the classical loop the human force F_h is the input to the fixed law [Equation (1)], whose output is the commanded correction x_d; the DRL loop replaces that fixed map with a policy $\pi\theta$ updated from the interaction reward. DRL: Deep reinforcement learning; HRI: human-robot interaction.

than holding a fixed mechanical law[15], so the mapping being learned is genuinely dynamic rather than merely unknown.

• **Non-stationarity.** Human stiffness, attention, and intent drift within and across sessions. Model-free policies can be updated online as this drift occurs, whereas fixed-gain admittance laws [Equation (1)] cannot.

• **Contact uncertainty.** Under unknown or time-varying environment constraints, residual strategies that learn a correction on top of a nominal force controller[4] and demonstration-guided exploration[16,17] have been shown to recover stable behaviour faster than learning from scratch.

Figure 1 illustrates this shift schematically: the classical loop (dashed) computes x_d directly from F_h via Equation (1); the DRL loop (solid) instead routes (s_t, F_h) through a learned policy π_θ, whose parameters are updated from the interaction reward, closing the loop at the learning level rather than only at the control level.

3. DEEP RL APPROACHES FOR HRI

We organise the DRL-for-HRI literature into four sub-categories, summarised in Table 2 and illustrated in Figure 2, that differ in what is being learned (a control law, a policy from data, a joint human-robot strategy, or a certified policy) rather than in application domain. The four categories build on the general DRL-for-robotics methodology established by policy-search and end-to-end visuomotor formulations[31,32], which they specialise to the case of a human in the loop.

The four categories are not disjoint, and their overlaps are significant for design. Demonstration learning (Section 3.2) and shared autonomy (Section 3.3) differ mainly in when the human's influence enters: in the former the human supplies an offline dataset or a preference label that shapes r before deployment, whereas in the latter the human is an agent whose action a_t^h appears inside the value function at run time. The boundary blurs as soon as demonstrations are collected online, at which point a demonstration-learning system is a shared-autonomy system with a particular arbitration rule. Safe RL (Section 3.4) is better read as orthogonal to the other three than as a peer of them: it is a constraint layer that can be applied to any of

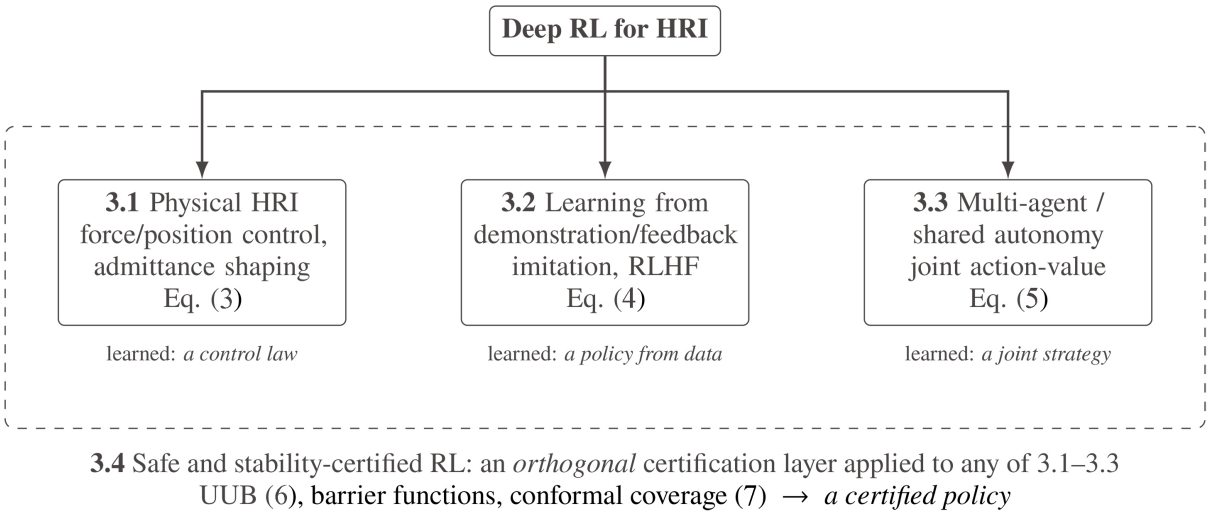


Figure 2. Classification of DRL approaches for HRI. Consistent with the 3+1 reading developed in the main text, categories 3.1-3.3 answer where human information enters and are drawn as siblings, whereas category 3.4 is not a fourth sibling but a certification layer (dashed band) that any of the other three can carry. Corresponds to Table 2. DRL: Deep reinforcement learning; HRI: human-robot interaction.

Table 2. Classification of DRL approaches for HRI

Category	Representative works	Learning mechanism	Typical HRI task
3.1 Physical HRI (force/position, admittance shaping)	[3,4,12-15]	RL-tuned virtual impedance/admittance gains, Equation (3)	Cooperative manipulation, physical assistance
3.2 Learning from demonstration/feedback	[16-18]	Imitation/behaviour cloning, RLHF preference models, Equation (4)	Sample-efficient policy initialisation from expert or preference data
3.3 Multi-agent/shared-autonomy RL	[19-22]	Two-agent Markov game, joint action-value function, Equation (5)	Redundant-manipulator co-manipulation, shared control
3.4 Safe and stability-certified RL	[7,8,23-30]	Lyapunov/UUB constraints, barrier functions, conformal coverage, Equations (6) and (7)	Safety-critical physical contact, certified deployment

DRL: Deep reinforcement learning; HRI: human-robot interaction; RL: reinforcement learning; UUB: uniform ultimate boundedness.

them. Concretely, demonstration data is what makes a safety filter tractable, because it supplies the calibration set that a conformal bound [Equation (7)] needs[28], while shared autonomy is what makes a safety filter necessary, because a non-stationary human partner is precisely the disturbance that a fixed-margin barrier cannot absorb[29]. Reading the four categories as a 3+1 structure, three answers to "where does human information enter" plus one certification layer, explains why the most deployment-ready systems in Table 3 combine categories rather than choosing between them.

Three observations follow from Table 3.

Deployment readiness runs inverse to conceptual ambition. The category that assumes least about the human, admittance shaping, is the one actually running on industrial hardware, while the category that models the human most richly, shared autonomy, has produced almost no deployed systems. This is not a transient state of maturity. Modelling the human as a rational agent buys expressiveness at the cost of an assumption that cannot be validated on the small cohorts these studies use, so the added realism is difficult to falsify in practice. In selecting a method for a given application, the assumptions listed in the second column are therefore the appropriate starting point.

Table 3. Critical comparison of the four categories: what each assumes, what it costs in data, and how close it is to deployment

Category	Core assumption	Principal limitation	Data requirement	Deployment readiness	Supporting evidence
3.1 Admittance shaping	Interaction is well described by a low-order mechanical law whose gains are the only unknowns[1,12]	Structure fixed at design time; cannot represent intent, only impedance; humans co-adapt force and impedance rather than holding a fixed law	Low: minutes of contact data, single user	High. Runs on stock force-controlled arms; the closest to industrial use	[1,3,12-15]
3.2 Demonstration/feedback	Expert or preference data is available, and the demonstrator's objective is recoverable	Inherits demonstrator bias and covariate shift; preference models are miscalibrated off-distribution	Moderate to high: 10^2-10^4 labelled comparisons or trajectories	Moderate. Strong in lab settings; annotation cost dominates at scale	[16-18]
3.3 Multi-agent/shared autonomy	The human is a rational agent with a stable, inferable policy	Human rationality and stationarity both fail; equilibrium concepts are hard to verify empirically	High: paired HRI logs, hard to parallelise	Low. Mostly simulation and small-cohort studies	[19-22]
3.4 Safe/certified RL	Disturbances are bounded, or calibration data is exchangeable with deployment data	Guarantees are conditional on assumptions the deployed system violates; conservatism costs task performance	Low for the certificate, but it inherits the data cost of whatever policy it wraps	Moderate and rising. Barrier-based filters ship; UUB-certified learned policies do not	[7,23-27,29]

HRI: Human-robot interaction; RL: reinforcement learning; UUB: uniform ultimate boundedness.

The assumptions fail together, not independently. Each row's assumption is violated by the same underlying fact, that the human adapts. Non-stationarity invalidates the fixed gains of 3.1, shifts the demonstration distribution of 3.2, breaks the stationary-policy premise of 3.3, and destroys the exchangeability that the conformal certificates of 3.4 require. Combining categories therefore does not average the risk away, because the failures are correlated: a system that pairs offline demonstrations with a conformal safety filter has two components that become miscalibrated at the same moment and for the same reason. Reported robustness gains from combining methods should be read with this in mind.

The data requirements in column four are not commensurable, which makes published comparisons unreliable. A demonstration costs human time, a preference label costs human attention, and a simulated rollout costs neither. Sample counts are nonetheless reported in a single currency, so a method requiring 10^4 simulated episodes is routinely described as less data-hungry than one requiring 10^2 human comparisons, when for a rehabilitation patient the ordering is reversed. Until the field reports human-cost separately from environment-interaction count, claims of improved sample efficiency in HRI are not comparable across papers. We regard this as a more tractable near-term fix than any algorithmic change, and as a concrete instance of the benchmarking challenge discussed in Section 5.

3.1. Physical HRI: force/position control and admittance shaping via RL

Rather than fixing M, D, K in Equation (1) offline, RL lets these parameters, or the resulting reference correction directly, be produced by a learned function of the interaction state,

$$x_d(t) = f_\theta\big(F_h(t), s_t\big), \qquad \theta \leftarrow \theta + \alpha\, \nabla_\theta J(\theta), \tag{3}$$

where $J(\theta)$ is the RL objective of Equation (2). The literature splits on how much structure to retain in f_θ. Variable-admittance schemes keep the mass-damper form and modulate D online from measured interaction power, which preserves a passivity argument at the cost of expressiveness[12,13]; variable-impedance learning instead treats the full stiffness profile as the learned object[3,14], which is more general but

forfeits the passivity certificate. Human motor control supplies the argument for the latter: humans co-adapt force and impedance and do not hold a fixed mechanical law[15], so a policy restricted to gain scheduling cannot reproduce the behaviour it is imitating. The trade-off is therefore not tuning *vs.* learning but guarantee *vs.* fidelity, and Section 3.4 is where that tension is addressed rather than resolved. When the contact model itself is unknown, residual RL[4] augments a nominal model-based term with a learned correction, anticipating the residual formulation discussed in Section 4.

3.2. Learning from human demonstration and feedback

Training π_θ from environment reward alone is sample-inefficient and can be unsafe during early exploration in physical contact with a human. Two complementary remedies dominate the literature we surveyed[16,17]: (i) imitation-style initialisation, where π_θ is pre-trained on expert trajectories before RL fine-tuning[17]; and (ii) preference-based reward shaping in the style of reinforcement learning from human feedback (RLHF)[18], where a reward model $\hat{r}_\phi$ is fit to human comparisons ($\tau^1 \succ \tau^2$) via a Bradley–Terry model,

$$P(\tau^1 \succ \tau^2) = \frac{\exp\left(\hat{r}_\phi(\tau^1)\right)}{\exp\left(\hat{r}_\phi(\tau^1)\right) + \exp\left(\hat{r}_\phi(\tau^2)\right)}, \tag{4}$$

and $\hat{r}_\phi$ then substitutes for, or augments, $r(s_t, a_t)$ in Equation (2). Safe-exploration variants constrain policy updates during this phase so that early, poorly-fit $\hat{r}_\phi$ or π_θ cannot command unsafe contact forces[23,25].

Which mechanism to use, and what each costs in HRI. These three mechanisms are often presented as interchangeable, and they are not. Behaviour cloning is a supervised fit to (s, a) pairs[16]: cheap, stable, and requiring no reward function, but it inherits the demonstrator's covariate shift, and in contact tasks that shift is unusually damaging, because a small deviation in commanded position becomes a large deviation in contact force through the environment stiffness. A cloned policy therefore degrades discontinuously rather than gracefully at the moment of contact. Inverse-RL-style imitation recovers a reward rather than a policy[17] and so extrapolates better off-distribution, but it requires solving a nested optimisation and assumes the demonstrator was itself optimal, an assumption that a human teaching through a force sensor, with fatigue and reaction delay, does not satisfy. Preference-based reward learning [Equation (4)][18] sidesteps the optimality assumption entirely, because a human need only compare, not perform, which is decisive for HRI: comparison is feasible for users who cannot demonstrate at all, such as rehabilitation patients. Its cost is sample complexity in the scarcest currency available, human attention, and a reward model that is confidently wrong outside the comparison distribution.

The HRI-specific difficulty that none of the three resolves is that the human is inside the loop being learned. In standard RLHF the annotator is external to the environment; in physical HRI the person supplying the preference is also the impedance the policy must control, so $\hat{r}_\phi$ and the transition P drift together. A policy that improves according to a fixed $\hat{r}_\phi$ may simply have adapted to a human who has adapted to it, which is not the same as improving. This coupling, rather than annotation cost, is in our view the binding constraint on preference-based methods in physical interaction, and it is not addressed by any of the works surveyed here.

3.3. Multi-agent and shared-autonomy RL

A second way to remove the fixed-model assumption is to stop treating the human as part of a stationary environment and instead model the interaction as a two-agent Markov game with joint action-value function

$$Q(s_t, a_t^h, a_t^r) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}^h, a_{t+k}^r) \,\middle|\, s_t, a_t^h, a_t^r\right], \tag{5}$$

where a_t^h and a_t^r are the human's and robot's actions, respectively. The field is divided according to what is inferred about the human. Policy blending interpolates between human and autonomous commands with an arbitration weight[20], which is simple and predictable but assists least exactly when the user's goal is most ambiguous. Hindsight optimisation removes that failure by treating the goal as a POMDP belief and assisting over the whole distribution[19], at the cost of a model of how the user acts. Deep RL formulations discard the goal model entirely and learn assistance end-to-end from interaction[21], buying generality and paying in sample complexity, which is the scarcest resource here [Table 3]. A broader review of intent detection, arbitration and feedback in physical shared control situates these choices[22]. What unites the category, and distinguishes it from the single-agent formulations of Section 3.1, is that the robot optimises its response to the human's ongoing action rather than to a fixed disturbance model.

3.4. Safe and stability-certified RL for HRI

This category asks not only what policy is learned but what can be guaranteed about it. UUB results for neural H_2 control[7] establish that tracking error e_t remains bounded,

$$\|e_t\| \le \varepsilon \quad \text{for all } t \ge T, \tag{6}$$

for some finite T and bound ε depending on approximation and disturbance terms, typically proved via a Lyapunov function $V(e_t)$ with $\dot{V}(e_t) < 0$ outside a compact set. Barrier-function methods[25] enforce a related but distinct guarantee, forward invariance of a safe set, directly inside the RL update. A complementary, distribution-free guarantee comes from conformal prediction[8]: given a calibration set, one constructs a prediction set $C(x)$ for an uncertain quantity (e.g., human motion or force) satisfying

$$P\big(y \in C(x)\big) \ge 1 - \alpha, \tag{7}$$

for a user-chosen miscoverage rate α, independent of the underlying predictor.

Conformal guarantees are already being used in HRI. A distinct line of work now couples Equation (7) to control-theoretic safety constraints in human-facing settings. Thumm *et al.* attach conformal prediction sets to vision-based human pose estimation and motion forecasting, propagate the resulting uncertainty end-to-end, and feed the sets into a certifiable safety framework validated on a physical human-robot collaboration cell; they also handle out-of-distribution inputs explicitly, which is the failure mode that defeats naive aleatoric uncertainty estimates[26]. Zhou *et al.* use adaptive conformal prediction to quantify motion-prediction uncertainty online and convert the resulting sets into probabilistic control-barrier-function constraints, so that the enforced safety margin adapts to the observed prediction error without assuming a noise distribution[27]. Gonzales *et al.* go further and apply conformal risk control to the control barrier function (CBF) safety value itself, tuning the safety margin online as a function of interaction context[28]. Busellato *et al.* fuse probabilistic hand-motion forecasting with CBFs so that the margin contracts when the forecast is confident, directly attacking the over-conservatism that makes worst-case envelopes unusable in practice[29].

The combination is therefore not absent, but asymmetric. Every one of these works pairs a distribution-free coverage guarantee [Equation (7)] with a set-invariance guarantee, that is, with barrier functions or predictive safety filters[23-25], because forward invariance composes naturally with a prediction set: the set enters the constraint as a margin. What remains genuinely underexplored is the pairing with stability guarantees of the UUB form [Equation (6)]. The reason is structural rather than accidental. A conformal set is a statement about a finite calibration sample and is exchangeability-dependent; a Lyapunov argument is a statement about a trajectory of a closed-loop system. Converting $P(y \in C(x)) \ge 1 - \alpha$ into a bound on the disturbance term that appears in $\dot{V}(e_t)$ requires an assumption about how coverage failures are distributed

over time, and the closed loop breaks exchangeability by construction, since the robot's own action changes the human's next motion. Adaptive conformal prediction[27] sidesteps this by abandoning the fixed-calibration assumption; a Lyapunov-based counterpart has no equivalent escape. Bridging that gap, producing a UUB bound whose disturbance term is certified by a conformal, closed-loop-valid argument, is the specific open problem this survey identifies.

Four routes to a combined certificate. We set out below how this gap might be closed. Write the closed-loop error dynamics with the human's motion or force entering as a disturbance d_t, so that a Lyapunov argument gives $\dot{V}(e_t) \leq -\kappa V(e_t) + \beta\|d_t\|$ and hence a bound ε in Equation (6) that is monotone in $\sup_t\|d_t\|$. The question is how to certify that supremum from data.

(i) Prediction set as a disturbance bound. The most direct route takes $C(x)$ from Equation (7) and reads off $\|d_t\| \leq \bar{d}(\alpha)$, giving a probabilistic UUB $\|e_t\| \leq \varepsilon(\alpha)$. The obstacle is that Equation (7) is a marginal statement about a single query, whereas $\sup_t$ requires coverage to hold simultaneously along a trajectory. A union bound over an H-step horizon degrades the confidence to $1 - H\alpha$, which is vacuous for the horizons of interest. This is the same difficulty solved in the planning literature by constructing prediction regions over whole predicted trajectories rather than per-step[36], and that construction transfers directly to the disturbance-bound setting.

(ii) Time-uniform coverage. A cleaner route replaces the fixed α with an anytime-valid statement[8], so that coverage holds simultaneously for all t without a union-bound penalty. This is the correct probabilistic object to pair with a Lyapunov argument, since both are trajectory-level statements, and to our knowledge it has not been attempted in HRI.

(iii) Average rather than pointwise coverage. Adaptive conformal prediction guarantees long-run empirical coverage without requiring exchangeability[27]. It therefore pairs naturally not with UUB but with an input-to-state-stability argument, in which the bound depends on an averaged disturbance norm rather than its supremum. Reformulating the guarantee from UUB to input to state stability (ISS) may be the least demanding route to a certified stability statement in closed loop, at the cost of a weaker conclusion.

(iv) Interventional calibration. Exchangeability of the calibration sample is what Equation (7) requires[8], and it fails here because the robot's action influences the human's next motion. It can be restored by construction if the calibration trajectories are collected under deliberately randomised robot actions, making the calibration distribution interventional rather than observational. The cost is that the randomisation must itself be safe, which returns the design to a barrier-based filter during data collection, and the practical question is how small the excitation can be while still breaking the correlation.

Assume-guarantee composition cuts across all four: certify the perception and prediction module conformally, certify the controller by Lyapunov argument, and connect them by an interface contract stating the disturbance bound each side assumes and provides. Conformal certification of learned perception feeding a control-theoretic guarantee is already established outside HRI[36], and within HRI the components exist separately, conformal perception on one side[26] and Lyapunov-plus-barrier control on the other[30]; what is missing is the contract that joins them. We regard route (ii) as the most promising theoretically and route (iii) as the most likely to appear first in practice.

4. EMERGING DIRECTIONS

Three directions extend the classification of Section 3 beyond what was addressed in[6] and are summarised in Table 4.

Table 4. Emerging directions in DRL for HRI

Direction	Ref.	Description
LLM/foundation-model-conditioned policies	[9,10]	Natural-language instructions are grounded into manipulation actions or task/motion plans, extending HRI from fixed tasks to open-vocabulary instruction following
Digital-twin-based sim-to-real transfer	[33,37]	A simulated replica of robot, human, and contact dynamics is used to pre-train π_θ safely before deployment, narrowing the sim-to-real gap identified as a persistent bottleneck for real-world DRL
Residual RL for hybrid model-based/model-free control	[4,11]	A learned correction Δa_t is added to a nominal model-based controller, Equation (8), combining the guarantees of Section 3.4 with the flexibility of Section 3.1

DRL: Deep reinforcement learning; HRI: human-robot interaction; LLM: large language model; RL: reinforcement learning.

Large language model (LLM)/Vision-language-action (VLA)-conditioned policies. Vision-language-action models such as RT-2[9] condition a manipulation policy directly on a natural-language instruction and an image observation, $\pi_\theta(a_t \mid s_t, \ell)$, where ℓ is the instruction embedding. Combined with task-and-motion planning[10], this allows an HRI system to accept unstructured verbal requests rather than a single fixed reward function, a qualitative departure from every method in Section 3, all of which assume a task-specific r fixed at training time. The trajectory here is not unique to robotics: medical vision-language analysis has passed through the same three stages, from task-specific models, through adapter- and prompt-tuned variants, to generalist foundation models, and the reported gains there are in data efficiency and cross-domain generalisation rather than in peak single-task accuracy[38]. That parallel is worth taking seriously as a prediction: if manipulation follows the same curve, the argument for VLA-conditioned policies in HRI will rest on few-shot adaptation to a new user, not on outperforming a well-tuned task-specific controller on the task it was tuned for.

Digital-twin-based sim-to-real transfer. Real-world DRL training in physical contact with a human is costly and risky; a digital twin of the human-robot-environment system allows π_θ to be pre-trained under Equation (2) in simulation, with the safety mechanisms of Section 3.4 used to certify the policy before, rather than during, real-world deployment. The transfer techniques this relies on, domain randomisation, system identification, and progressive adaptation, are surveyed in[37]; surveys of real-world DRL deployment in robotics consistently flag the residual sim-to-real gap in human-facing settings as an open bottleneck[33], motivating this direction as a priority rather than a solved problem.

4.1. Data-efficient and offline learning for HRI

The methods of Section 3 share a dependence that Table 3 makes explicit: they need interaction data, and in HRI every sample is a physical interaction with a person. This section covers the four bodies of work that attack that constraint directly.

Offline RL is the structurally indicated response, not an optional extension. If human-in-the-loop rollouts cannot be reset or parallelised, then learning must proceed from a fixed dataset of previously logged interactions; this is precisely the offline RL setting[39,40]. The central difficulty there, distributional shift, is that a policy trained on logged data queries the value function at actions the behaviour policy never took, and the resulting overestimation is unrecoverable without online correction. Conservative and constraint-based estimators address this by penalising out-of-distribution actions[40]. HRI aggravates the problem along an axis the offline RL literature does not treat: the shift is not only between behaviour policy and learned policy, but between the logged human and the deployed human. A dataset collected from one cohort encodes that cohort's stiffness, reaction latency, and intent distribution, so an offline HRI policy is conservative with respect to the wrong reference. This is the generalisation-across-users challenge of Section 5 restated in

offline terms, and it suggests that per-user conservatism, rather than per-action conservatism, is the appropriate regulariser. We are not aware of an offline RL method formulated this way.

Data efficiency beyond offline learning. Three of the mechanisms already surveyed are data-efficiency mechanisms in disguise: demonstrations and preferences replace environment interaction with human supervision (Section 3.2), sim-to-real transfer replaces it with simulated interaction[37], and residual RL replaces it with a nominal controller that is already approximately correct [Equation (8)]. Their limitations are complementary rather than shared, which is why they compose: imitation fails off-distribution, simulation fails where contact dynamics are mismodelled, and residual formulations fail when the nominal controller is badly wrong. What has been missing is measurement. Benchmarks that evaluate offline methods on real robot hardware rather than in simulation[41] report substantially smaller gains than simulated benchmarks suggest, which is a caution worth carrying into HRI, where the gap between simulated and real interaction is wider still because the human is the least faithfully simulated component.

Foundation models change what counts as a data requirement. The VLA policies of Section 4 are trained on cross-embodiment corpora aggregated across many robots and laboratories[42], so the marginal cost of a new task shifts from collecting a task-specific dataset to specifying an instruction. For HRI, the interesting claim is not peak task performance but few-shot adaptation to a new user, which is the axis on which the methods of Section 3 generalise worst. The same three-stage progression, from task-specific models through adapter-tuned variants to generalist foundation models, has already played out in medical vision-language analysis, where the reported gains were likewise in data efficiency and cross-domain transfer rather than in single-task accuracy[38]. Whether physical interaction follows that curve is open: contact dynamics are not obviously compressible in the way that visual semantics are.

Uncertainty-aware learning. The fourth strand is treated at length in Section 3.4, where conformal prediction sets and their composition with barrier functions and predictive safety filters are surveyed[26-30]. We note here only the connection to the present section: a conformal guarantee is a statement about a calibration set, so it inherits exactly the coverage limitations of offline data discussed above. An offline HRI policy and its conformal certificate can be miscalibrated for the same reason and at the same time, which is a failure mode that neither literature currently isolates.

Residual RL for hybrid control. Rather than learning π_θ from scratch, residual RL composes a nominal model-based action a_t^{model} {e.g., an admittance law [Equation (1)] or an inverse-dynamics controller} with a learned correction,

$$a_t = a_t^{\text{model}} + \Delta a_t^{\text{RL}}(s_t), \qquad \Delta a_t^{\text{RL}} \sim \pi_\theta(\cdot \mid s_t), \tag{8}$$

which preserves the nominal controller's stability properties when Δa_t^{RL} is small, while still allowing the policy to compensate for unmodeled human or contact dynamics. Figure 3 shows the resulting architecture, in which the online constraint [Equation (7)] is applied to the composed action before it reaches the plant. The formulation was introduced for contact-rich robot control in[4,11], and we expect it to be the most practical bridge between the certified guarantees of Section 3.4 and the flexibility of end-to-end DRL.

5. OPEN CHALLENGES

Despite the progress organised in Tables 1 and 2, five challenges remain largely open across the DRL-for-HRI literature we surveyed. Table 5 summarises them; each is discussed below.

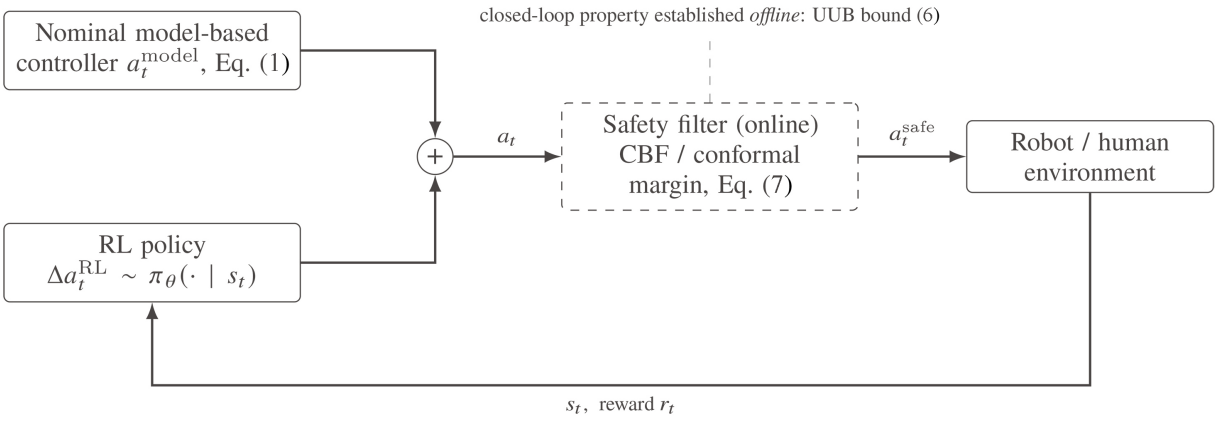


Figure 3. Residual RL architecture [Equation (8)], combining a nominal model-based controller with a learned correction. Only the set-invariance/conformal constraint [Equation (7)] is executable as an online filter that modifies a_t; the UUB bound [Equation (6)] is a property of the resulting closed loop established by analysis, not a block in the signal path, and is therefore shown separately. RL: Reinforcement learning; UUB: uniform ultimate boundedness.

Table 5. Open challenges in DRL for HRI

Challenge	Related work	Core difficulty
Sample efficiency with humans in the loop	[16,17]	Real human-in-the-loop rollouts are slow, fatiguing, and cannot be parallelised or reset like simulation, limiting the number of policy updates
Safety certification	[7,8,23,25,26,28]	Conformal coverage now composes with set invariance in HRI, but not with UUB stability [Equation (6)]: the closed loop breaks the exchangeability that conformal calibration assumes
Non-stationary human behaviour	[15,22]	A policy trained on one user's stiffness/intent profile can degrade as the human adapts, fatigues, or is replaced by a different user, violating the stationarity assumption behind Equation (2)
Generalisation across users	[12,22]	Most reported results are single-user or small-cohort; policies tuned to one person's dynamics do not reliably transfer without re-training or online adaptation
Benchmarking and reproducibility	[33,35]	HRI experiments vary in hardware, human subjects, and reward design, so results across papers are difficult to compare; no HRI-specific counterpart to standard DRL benchmarks has been widely adopted

DRL: Deep reinforcement learning; HRI: human-robot interaction; UUB: uniform ultimate boundedness.

Each challenge below is stated as a specific unresolved question rather than a topic, together with what would count as an answer. Several were surfaced by the comparative analysis of Section 3 and are, to our knowledge, not posed in this form elsewhere.

C1. Is per-user conservatism the right regulariser for offline HRI? Offline RL penalises actions outside the behaviour distribution[39,40], but Section 4.1 argued that the shift that matters in HRI is between the logged human and the deployed human, not between behaviour and learned policy. The concrete question: can a conservatism penalty be defined over an estimated user-parameter posterior, for instance over arm stiffness and reaction latency, rather than over actions? An answer would be an offline method whose pessimism relaxes as the deployed user's estimated parameters approach the logged cohort, and it could be evaluated today on existing real-robot offline benchmarks[41] extended with multi-user data.

C2. Can conformal coverage be made valid in closed loop? Section 3.4 showed that conformal sets now compose with set-invariance guarantees in HRI[26-29] but not with Lyapunov/UUB stability, because the robot's action changes the human's next motion and so destroys exchangeability. The specific open problem is to bound the disturbance term in $\dot{V}(e_t)$ using a coverage statement that remains valid under this feedback. Adaptive conformal prediction[27] shows one route, abandoning fixed calibration; a second, untried route is

to calibrate on interventional rather than observational data, that is, on trajectories in which the robot's action was deliberately randomised. The answer would be a UUB bound whose ε in Equation (6) carries an explicit, finite-sample confidence level.

C3. Do the four categories fail independently? The synthesis following Table 3 argued they do not: human adaptation invalidates the assumptions of all four simultaneously. If correct, this predicts that a system pairing offline demonstrations with a conformal safety filter will show correlated degradation, both components decalibrating within the same session. This is directly testable, and to our knowledge untested: instrument a combined system, log calibration error and policy regret separately, and measure their correlation across a session in which the user adapts. A near-zero correlation would falsify the claim and justify treating the categories as independent risk reducers; a high correlation would mean that published robustness gains from combining methods are overstated.

C4. How many users, of what diversity, are enough? Most reported results are single-user or small-cohort[12,22]. The unresolved question is not qualitative but quantitative: what is the sample complexity of user generalisation? A concrete first step is a scaling study holding the algorithm fixed and varying cohort size and diversity, reporting held-out-user performance as a curve rather than a single number. Without such a curve there is no principled basis for the cohort sizes currently used, which appear to be set by recruitment convenience.

C5. Report human cost separately from interaction count. As argued after Table 3, demonstrations, preference labels and simulated rollouts are aggregated into a single sample count even though they draw on incommensurable resources, so a method needing 10^4 simulated episodes is described as more efficient than one needing 10^2 human comparisons. This item requires no new algorithm. We propose that HRI-DRL papers report a three-element cost vector: environment interactions, human minutes, and number of distinct participants. Unlike a full benchmark suite, which requires community coordination[33,35], this convention can be adopted by an individual paper without prior agreement across the field.

6. CONCLUSION

This survey organised DRL for HRI into four paradigms, physical HRI via admittance shaping, learning from demonstration and feedback, multi-agent and shared-autonomy control, and safe, stability-certified RL, and traced their evolution from the fixed-law control reviewed in our 2021 book[6] to the certified, language-conditioned, and residual formulations emerging today. Relative to recent DRL-robotics surveys in which HRI appears as one application area among many[33], and to surveys addressing a specific HRI setting or a specific communication channel rather than control-theoretic guarantees[34,35], we focused specifically on what a robot can be guaranteed to do while interacting with a human. Section 3.4 argues that the residual weakness is not the absence of distribution-free uncertainty quantification in HRI, which now exists[26-29], but its confinement to set-invariance rather than stability guarantees.

Three directions are, in our view, the most consequential for closing that gap: reconciling conformal-prediction coverage[8], which is now routinely composed with set-invariance guarantees in HRI[26-29], with Lyapunov/UUB stability results[7] into a single certification framework, which requires a coverage argument that survives the loss of exchangeability in closed loop; using digital twins to move safety validation from real-world trials into simulation before deployment; and adopting residual RL [Equation (8)] as a default architecture, since it inherits the guarantees of a nominal controller while retaining the adaptability that motivated moving beyond classical admittance control [Equation (1)] in the first place. We hope the classification and open challenges identified here [Tables 2 and 5] provide a concrete agenda for both the RL and HRI communities.

DECLARATIONS

Authors' contributions

The author contributed solely to the article.

Availability of data and materials

Not applicable.

AI and AI-assisted tools statement

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

Yu, W. is a Section Member of the Section *Human-Computer Interaction, Language, and Collaboration* of *Intelligence & Robotics*. Yu, W. was not involved in any step of the editorial process for this manuscript, including reviewer selection, manuscript handling, or the editorial decision.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Hogan, N. Impedance control: an approach to manipulation: Part I - theory. *J. Dyn. Sys. Meas. Control.* **1985**, *107*, 1-7. DOI
2. Billard, A.; Kragic, D. Trends and challenges in robot manipulation. *Science* **2019**, *364*, eaat8414. DOI PubMed
3. Abu-Dakka, F. J.; Saveriano, M. Variable impedance control and learning - a review. *Front. Robot. AI.* **2020**, *7*, 590681. DOI PubMed PMC
4. Johannink, T.; Bahl, S.; Nair, A.; et al. Residual reinforcement learning for robot control. In *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, Canada. May 20-24, 2019. IEEE; 2019. pp. 60239. DOI
5. Mir, B. A.; Nishwa, D. E.; Lee, S. W. Embodied artificial intelligence in healthcare: a systematic review of robotic perception, decision-making, and clinical impact. *Healthcare* **2026**, *14*, 572. DOI PubMed PMC
6. Yu, W.; Perrusquía, A. *Human-robot interaction control using reinforcement learning*. Hoboken, NJ, USA: Wiley-IEEE Press, 2021. DOI
7. Perrusquia, A.; Yu, W. Neural H_2 control using continuous-time reinforcement learning. *IEEE. Trans. Cybern.* **2022**, *52*, 4485-94. DOI PubMed
8. Angelopoulos, A. N.; Bates, S. Conformal prediction: a gentle introduction. *Found. Trends. Mach. Learn.* **2023**, *16*, 494-591. DOI
9. Brohan, A.; Brown, N.; Carbajal, J.; et al. RT-2: vision-language-action models transfer web knowledge to robotic control. *arXiv* **2023**, arXiv:2307.15818. Available online: https://doi.org/10.48550/arXiv.2307.15818. (accessed on 2026-08-21).
10. Ding, Y.; Zhang, X.; Paxton, C.; Zhang, S. Task and motion planning with large language models for object rearrangement. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Detroit, USA. Oct 01-05, 2023. IEEE; 2023. pp. 2086-92. DOI
11. Silver, T.; Allen, K.; Tenenbaum, J.; Kaelbling, L. Residual policy learning. *arXiv* **2018**, arXiv:1812.06298. Available online: https://doi.org/10.48550/arXiv.1812.06298. (accessed on 2026-08-21).
12. Keemink, A. Q. L.; van der Kooij, H.; Stienen, A. H. A. Admittance control for physical human–robot interaction. *Int. J. Robot. Res.* **2018**, *37*, 1421-44. DOI
13. Ferraguti, F.; Talignani Landi, C.; Sabattini, L.; Bonfè, M.; Fantuzzi, C.; Secchi, C. A variable admittance control strategy for stable physical human–robot interaction. *Int. J. Robot. Res.* **2019**, *38*, 747-65. DOI
14. Ficuciello, F.; Villani, L.; Siciliano, B. Variable impedance control of redundant manipulators for intuitive human-robot physical interaction. *IEEE. Trans. Robot.* **2015**, *31*, 850-63. DOI

15. Yang, C.; Ganesh, G.; Haddadin, S.; Parusel, S.; Albu-Schaeffer, A.; Burdet, E. Human-like adaptation of force and impedance in stable and unstable interactions. *IEEE. Trans. Robot.* **2011**, *27*, 918-30. DOI
16. Argall, B. D.; Chernova, S.; Veloso, M.; Browning, B. A survey of robot learning from demonstration. *Robot. Auton. Syst.* **2009**, *57*, 469-83. DOI
17. Ravichandar, H.; Polydoros, A. S.; Chernova, S.; Billard, A. Recent advances in robot learning from demonstration. *Annu. Rev. Control. Robot. Auton. Syst.* **2020**, *3*, 297-330. DOI
18. Kaufmann, T.; Weng, P.; Bengs, V.; Hüllermeier, E. A survey of reinforcement learning from human feedback. *arXiv* **2023**, arXiv:2312.14925. Available online: https://doi.org/10.48550/arXiv.2312.14925. (accessed on 2026-08-21).
19. Javdani, S.; Admoni, H.; Pellegrinelli, S.; Srinivasa, S. S.; Bagnell, J. A. Shared autonomy via hindsight optimization for teleoperation and teaming. *J. Robot. Res.* **2018**, *37*, 717-42. DOI
20. Dragan, A. D.; Srinivasa, S. S. A policy-blending formalism for shared control. *Int. J. Robot. Res.* **2013**, *32*, 790-805. DOI
21. Reddy, S.; Dragan, A. D.; Levine, S. Shared autonomy via deep reinforcement learning. *arXiv* **2018**, arXiv:1802.01744. Available online: https://doi.org/10.48550/arXiv.1802.01744. (accessed on 2026-08-21).
22. Losey, D. P.; Mcdonald, C. G.; Battaglia, E.; O'Malley, M. K. A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction. *Appl. Mech. Rev.* **2018**, *70*, 010804. DOI
23. Brunke, L.; Greeff, M.; Hall, A. W.; et al. Safe learning in robotics: from learning-based control to safe reinforcement learning. *Annu. Rev. Control. Robot. Auton. Syst.* **2022**, *5*, 411-44. DOI
24. Hewing, L.; Wabersich, K. P.; Menner, M.; Zeilinger, M. N. Learning-based model predictive control: toward safe learning in control. *Annu. Rev. Control. Robot. Auton. Syst.* **2020**, *3*, 269-96. DOI
25. Cheng, R.; Orosz, G.; Murray, R. M.; Burdick, J. W. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. *arXiv* **2019**, arXiv:1903.08792. Available online: https://doi.org/10.48550/arXiv.1903.08792. (accessed on 2026-08-21).
26. Thumm, J.; Frei, M.; Ni, T.; Althoff, M.; Pavone, M. Vision-based safe human-robot collaboration with uncertainty guarantees. *arXiv* **2026**, arXiv:2604.15221. Available online: https://doi.org/10.48550/arXiv.2604.15221. (accessed on 2026-08-21).
27. Zhou, H.; Zhang, Y.; Luo, W. Safety-critical control with uncertainty quantification using adaptive conformal prediction. In *2024 American Control Conference (ACC)*, Toronto, Canada. Jul 10-12, 2024. IEEE; 2024. pp. 574-80. DOI
28. Gonzales, J.; Mizuta, K.; Leung, K.; Ratliff, L. J. Safe probabilistic planning for human-robot interaction using conformal risk control. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hangzhou, China. Oct 19-25, 2025. IEEE; 2025. pp. 18676-83. DOI
29. Busellato, L.; Cunico, F.; Dall'Alba, D.; et al. Uncertainty aware-predictive control barrier functions: safer human–robot interaction through probabilistic motion forecasting. *Robot. Auton. Syst.* **2026**, *197*, 105291. DOI
30. Zhang, D.; Van, M.; McIlvanna, S.; Sun, Y.; McLoone, S. Adaptive safety-critical control with uncertainty estimation for human-robot collaboration. *IEEE. Trans. Autom. Sci. Eng.* **2024**, *21*, 5983-96. DOI
31. Kober, J.; Bagnell, J. A.; Peters, J. Reinforcement learning in robotics: a survey. *Int. J. Robot. Res.* **2013**, *32*, 1238-74. DOI
32. Levine, S.; Finn, C.; Darrell, T.; Abbeel, P. End-to-end training of deep visuomotor policies. *arXiv* **2015**, arXiv:1504.00702. Available online: https://doi.org/10.48550/arXiv.1504.00702. (accessed on 2026-08-21).
33. Tang, C.; Abbatematteo, B.; Hu, J.; Chandra, R.; Martín-Martín, R.; Stone, P. Deep reinforcement learning for robotics: a survey of real-world successes. *Annu. Rev. Control. Robot. Auton. Syst.* **2025**, *8*, 153-88. DOI
34. Eskue, N.; Baptista, M. L. Deep reinforcement learning for facilitating human-robot interaction in manufacturing. In: Islam, M. M. M.; Baptista, M. L.; Tariq, F.; editors. *Artificial intelligence for smart manufacturing and industry X.0*. Cham: Springer Nature Switzerland; 2025. pp. 69-95. DOI
35. Habibian, S.; Alvarez Valdivia, A.; Blumenschein, L. H.; Losey, D. P. A survey of communicating robot learning during human-robot interaction. *Int. J. Robot. Res.* **2025**, *44*, 665-98. DOI
36. Lindemann, L.; Cleaveland, M.; Shim, G.; Pappas, G. J. Safe planning in dynamic environments using conformal prediction. *IEEE. Robot. Autom. Lett.* **2023**, *8*, 5116-23. DOI
37. Zhao, W.; Peña Queralta, J.; Westerlund, T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Canberra, Australia. Dec 01-04, 2020. IEEE; 2020. pp. 737-44. DOI
38. Ali, M. U.; Zafar, A.; Kim, S.; Kim, K. S.; Lee, S. W. From task-specific to foundation models: a paradigm shift in medical vision-language analysis. *Comput. Sci. Rev.* **2026**, *59*, 100831. DOI
39. Levine, S.; Kumar, A.; Tucker, G.; Fu, J. Offline reinforcement learning: tutorial, review, and perspectives on open problems. *arXiv* **2020**, arXiv:2005.01643. Available online: https://doi.org/10.48550/arXiv.2005.01643. (accessed on 2026-08-21).
40. Figueiredo Prudencio, R.; Maximo, M. R. O. A.; Colombini, E. L. A survey on offline reinforcement learning: taxonomy, review, and open problems. *IEEE. Trans. Neural. Netw. Learn. Syst.* **2024**, *35*, 10237-57. DOI

41. Gürtler, N.; Blaes, S.; Kolev, P.; et al. Benchmarking offline reinforcement learning on real-robot hardware. *arXiv* **2023**, arXiv:2307.15690. Available online: https://doi.org/10.48550/arXiv.2307.15690. (accessed on 2026-08-21).
42. Open X-Embodiment Collaboration. Open X-embodiment: robotic learning datasets and RT-X models. *arXiv* **2023**, arXiv:2310.08864. Available online: https://doi.org/10.48550/arXiv.2310.08864. (accessed on 2026-08-21).