Human-Robotic Prosthesis as Collaborating Agents for Symmetrical Walking

Anonymous Author(s)
Affiliation
Address
email

Abstract

This is the first attempt at considering human influence in the reinforcement learning control of a robotic lower limb prosthesis toward symmetrical walking in real world situations. We propose a collaborative multi-agent reinforcement learning (cMARL) solution framework for this highly complex and challenging human-prosthesis collaboration (HPC) problem. The design of an automatic controller of the robot within the HPC context is based on accessible physical features or measurements that are known to affect walking performance. Comparisons are made with the current state-of-the-art robot control designs, which are single-agent based, as well as MARL solution approaches tailored to the problem, including multi-agent deep deterministic policy gradient (MADDPG) and counterfactual multi-agent policy gradient (COMA). Results show that, when compared to these approaches, treating human and robot as coupled agents and using estimated human adaption in robot control design accelerates learning of walking tasks and increases learning success rate. The proposed framework can potentially be further developed to examine how human and robotic lower limb prosthesis interact, an area that little is known about. Advancing cMARL toward real world applications such as HPC for normative walking sets a good example of how AI can positively impact on people's lives.

1 Introduction

The concept, design, and applications of human-robot cooperation have advanced rapidly due to new demands in AI-enabled applications fueled by powerful deep learning and reinforcement learning algorithms [1, 2, 3, 4]. Human-robot collaboration can take on a variety of forms depending on tasks to be solved, how information is shared, and the nature of interaction [2, 3, 4]. Examples of complex collaborative tasks include picking up or carrying objects together, cooperating on a production line, in which cases a robot can learn to imitate human demonstrations [5, 6, 7, 8, 9]. Other application scenarios may include intermittent robotic correction of human driving, or vice versa [10]. In essence, most of these recent studies involve interactions between a human and a robot in a way such that there is either space between the agents, or there is time for predictive counter measures to take place. Lower limb exoskeleton is another major type of wearable robot. The conceptualization of the control problem is different even though recent designs also take into account human-robot interacting effects [11, 12] as these exoskeletons mainly focus on end point performance of a foot or lower limb joints where user intent is quantified by estimating human joint torque or interactive torque. Thus, human-exoskeleton system performance goal is for the robot to produce a well-defined joint or endpoint trajectory. [13, 14, 15].

Seamless collaboration between a robotic lower limb prosthesis serially attached to its human user is critical to achieve normative walking [16, 17]. Yet, the problem is unique and it is challenging

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

to appropriately set up human-robot performance goals. The human-robot collaborative walking problem involves two strongly coupled agents, the interacting dynamics of which are difficult or nearly impossible to describe by an ordinary difference or differential equation. Additionally, human perception and cognition change when wearing a prosthesis [18]. Locomotion of a human-robot system may be affected by several factors, such as lower limb mechanics, inter-limb neuromechemical coupling, and physical structure of the human body including the lower limb [19, 20]. Physical and physiological differences in individual human subjects further complicates the robot control design. Even though our knowledge of human motor control and motor learning, in cases such as post stroke or a general loss of normative locomotion capabilities, have expanded greatly [21], little is known about how an active robotic prosthesis, not a traditional stick-type passive prosthesis, affects human and vice versa in walking tasks. This is because that the human-robotic prosthesis interacts continuously, the intricate human neurocontrol circuits including sensing, perception and feedback control, which are necessary to facilitate normative walking, is disrupted after amputation.

Innovative solution frameworks are needed to first address the design, analysis, and testing of complex control systems to restore locomotion for amputees, based on which, human-robot interactions can be studied, and fundamental issues such as user preference and prosthesis embodiment can be examined. The knowledge in turn, will further help design seamless automatic robot control systems. As such, challenges arising from controlling a wearable robotic lower limb prosthesis to meet the human user's needs are unique, and in some aspects, these challenges are greater than what have been studied [16, 17, 22, 23].

Achieving normative walking is fundamentally a real time control problem that involves continuous states and continuous controls of the human-robot system. State-of-the-art automatic control of robotic prosthesis has been reported recently with successful real human subject testing [24, 25]. These approaches are single agent-based reinforcement learning controls, specific tasks such as level ground and ramp walking were accomplished via designer prescribed or specified robotic joint movement profiles. A most recent single-agent based reinforcement learning solution demonstrated that human amputee subject can perform level ground and ramp walking with the robot controller aimed at mimicking the intact joint movement. Note, however, that all previous results have not considered human influence on the human-robot system performance in the automatic control of the robotic prosthesis, a phenomena that has direct impact not only on restoring walking but also on human health [26]. Multi-agent reinforcement learning control is a natural candidate to address this challenge. Yet, only a handful of results are available, and they are not directly in the problem domain under our consideration. For example, existing work on the physical human-exoskeleton collaboration treated the problem in a leader (human)-follower (exoskeleton) setting. Leader-follower coordination was achieved by measured/predicted interaction force [27]. Such an approach was to have the robot follow the learned behavior from human.

Great effort has been made develop continuous control reinforcement learning algorithms which can substantively address real world applications. Several algorithms, such as deep deterministic policy gradient (DDPG) [28], proximal policy optimization (PPO) [29], soft actor-critic (SAC) [30], and twin delayed DDPG (TD3) [31], have demonstrated success with solving complex control problems. For example, in simulated human locomotion control [32, 33], over 20 independent control signals are used to control a humanoid robot to achieve different walking tasks. Robotic upper-limb control has undergone intense development, especially in the realms of robot surgical assistance [34, 35], patient rehabilitation [34], and industrial applications [36]. However, these problems differ fundamentally from the human-robot walking problem, as their control target usually consists of well-defined end points generated by a decoupled trajectory generation exosystem [37]. On the other hand, walking tasks are difficult to associate with an end point task goal due to heavy dynamic coupling between the human reference and robotic limb. Single agent-based continuous control has also demonstrated promise in engineering applications such as stabilization, tracking, and reconfiguring control of Apache helicopters [38, 39, 40], stabilization and control of large power grids [41, 42, 43], robotic manipulation and locomotion via MuJoCo and OpenAI Gym [44, 45], and wearable robots with human in the loop [46, 47, 48]. While these works are encouraging toward solving realistic single-agent continuous control problems, it is not obvious how they can directly address the multi-agent human-robot normative walking problem.

2 Related work

Single agent RL for automatic control of a robotic limb. Most state-of-the-art RL control design approaches to enable continuous human-robot walking are single-agent based. Important milestones have been achieved by two major classes of RL algorithms: 1) actor-critic algorithms, including direct heuristic dynamic programming (dHDP) and Q-learning [46, 47, 25, 49, 50, 51], and 2) variants of policy iteration algorithms, including flexible policy iteration (FPI) [52] and policy iteration with constraint embedded (PICE) [53, 54]. Both types of single-agent based control algorithms were tested on amputee subjects walking with a robotic knee prosthesis. In human tests, 12 prosthesis control inputs were tuned within approximately 10 minutes of level ground walking. The tuned single-agent controllers were then tested on human subjects walking at different speeds and in different terrain conditions [53]. Note that all of the above methodologies and tests used designer-prescribed robot joint movement profiles generated *a priori* for the specific subjects and walking tasks. In real world use scenarios, joint movement profiles must evolve dynamically in real-time to accommodate internal human walking objectives. Additionally, profiles will vary significantly between individuals, even for execution of the same task in the same environment. A more recent single-agent RL control work replaced designer-prescribed joint movement profiles with the intact joint motion, achieving great success in human amputee tests . The current study, therefore, makes use of this idea as a constituent element of controller design. Note also that, none of the above results have taken into account human influence on human-robot walking in the control design, a novel contribution of this work.

Multi-agent reinforcement learning (MARL). A cMARL problem usually cannot be transformed to an equivalent single-agent problem, as additional agent(s) and their control policies introduce uncertainties and/or non-stationarity to the environment [55]. Partial observability is common in cMARL problems, which are usually addressed by the decentralized POMDP framework. Centralized training decentralized execution (CTDE) is a popular paradigm to address multi-agent coordination problems. A central feature of most CTDE approaches is factorization of the joint state-action value function into individual utility functions. **MADDPG** [56] is a popular CTDE-based method to solve both cooperative and competitive MARL problems. Agents are assumed to communicate complete and perfect information with each other, so a centralized critic is trained from global state/action data, allowing agents to have individual reward functions. However, since it is unrealistic for the human to have direct access to robot kinematic sensor data, this structure is not immediately appropriate for the human-robot collaborative walking problem. **COMA** [57] is another popular MARL algorithm which utilizes a single centralized critic by using global state information and actions of all agents. Additionally, COMA addresses the multi-agent credit assignment challenge by using a counterfactual baseline which keeps all other agents' actions fixed while treating a single marginalized agent. In game-playing problems, COMA significantly improves performance over other MARL methods such as independent Q-learning [58], multiagent bidirectionally-coordinated nets [59], and variant DQN [60]. However, due to the heavy dynamic coupling between the human and prosthesis, it is impossible to keep one agent action fixed and marginalize out the other agent's action. As a result, this counterfactual baseline is not appropriate for our problem, either.

Contributions. Human-robot collaborative tasks will continue to play an increasingly vital role in modern life, and existing single- and multi-agent control frameworks fail to substantively address these classes of problems. As such, the contributions of this work are as follows. 1) We introduce an innovative cMARL design approach to automatically control a robotic lower limb prosthesis by treating the human user as a collaborating agent whose control action is used to represent human influence on human-robot symmetrical walking performance. 2) We demonstrate that by having the prosthetic knee track the intact knee, the proposed framework does not require a control designer to prescribe a robot control target for each walking task and for each person, overcoming a fundamental design barrier facing existing methods in real life applications. 3) Based on three sets of simulation studies (benchmark, ablation, and evaluation), results show that the proposed design outperforms benchmark MARL methods tailored to the problem, as well as state-of-the-art single agent based control designs.

3 Fundamental control problem statement

In powered lower limb prosthesis, the finite state impedance control (FS-IC) framework most frequently serves to provide intrinsic control of a robotic joint, i.e., it is the built-in controller from

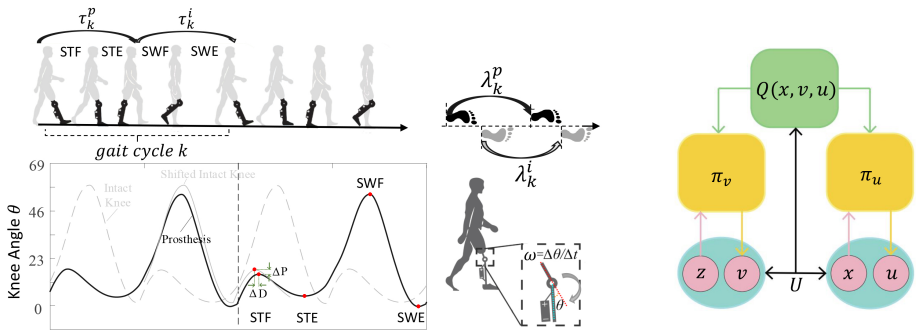

Figure 1: Human-robot as collaborating agents for symmetrical walking. Left: human state variables include step length difference and stance time difference that are also accessible to the robot, while the robot state includes two additional robot kinematic variables. Right: our cMARL solution approach to the symmetrical walking problem.

robotic prosthesis manufacturers. While position control is common in industrial robots, when a human is paired with a robotic joint, robot trajectory tracking using position control may preclude any dynamic interaction of the robot with its human user and the environment [61]. This may cause an amputee to react to the awkwardness of the prosthesis rather than to interact constructively with it [62]. As FS-IC impedance control framework is considered to provide compliant control for human use of the robot, it is expected to generate stable and predictable human-robot walking behavior, and thus, it is adopted in this study which aims for its controller designs at real world applications.

The FS-IC treats a gait cycle, or a step, as four consecutive gait phases (Figure 1, Left): stance flexion (STF) as Phase 1, stance extension (STE) as Phase 2, swing flexion (SWF) as Phase 3, and swing extension (SWE) as Phase 4. In what follows, we design four individual controllers for each of the four phases. Since the four controllers are conceptualized similarly and designed using the same approach, for the sake of clarity we carry out subsequent discussion without explicitly referring to the specific phase numbers (but we emphasize that each of the four phases requires its own independently-trained controller). Additional details on FS-IC such as phase detection and transition are provided in Appendix A.3,

At the k-th gait cycle, a robot controller (solved by cMARL in this paper) is to determine three impedance parameters, $I_k = [K_k, B_k, (\theta_e)_k]^T \in \mathbb{R}^3$ under the FS-IC framework, representing stiffness K_k, damping coefficient B_k, and equilibrium position $(\theta_e)_k$, respectively. The prosthetic joint motor torque $T_k \in \mathbb{R}$ is then generated based on joint kinematics (knee joint angle θ and angular velocity ω) according to the following impedance control law,

$$T_k = K_k(\theta - (\theta_{e_k})) + B_k\omega. \tag{1}$$

The control problem formulation requires automatically determining 12 control inputs or impedance control parameters (3 for each of the 4 gait phases) for individual users aiming at walk in real world situations. The initial set of feasible baseline impedance control parameters I_0 can be obtained from manufacturers or rehabilitation clinics. Impedance updates take place according to

$$I_{k+1} = I_k + u_k, \tag{2}$$

where $u_k \in \mathbb{R}^3$ is to be determined from our proposed cMARL approach.

4 Method

The human user and the robotic limb are considered collaborating agents. Our cMARL solution approach toward human-robot symmetrical walking problem is formulated based on physical features and measurements that have been shown affecting human-robot walking performance, and they are respectively available to the human and the robot, but not necessarily to each other.

State and control variables. Refer to Figure 1. At the k-th gait cycle, let τ_k^i and λ_k^i, respectively, represent the stance time (time of foot on the ground) and step length (length between two consecutive

steps when toe touching the ground) of the human intact leg. Similarly, τ_k^p and λ_k^p, respectively, for the prosthetic leg. Let $\Delta\tau_k$, $\Delta\lambda_k$ denote respectively the difference of stance time and step length,

$$\Delta\tau_k = \tau_k^p - \tau_k^i,\ \Delta\lambda_k = \lambda_k^p - \lambda_k^i. \tag{3}$$

We thus define human state variable $z_k \in \mathbb{R}^2$ as $z_k = [\Delta\tau_k, \Delta\lambda_k]^T$.

Robot states include kinematic features determined from the knee movement profiles as well as step length and stance time (or z_k). For robot kinematic state variables, we extract gait kinematic features from knee motion profiles of both limbs (Figure 1). Let P_k^i and D_k^i represent the knee angle and time duration of the intact limb, and similarly, P_k^p and D_k^p for the prosthetic limb. Let ΔP_k and ΔD_k denote the error of peak knee angles and the error of duration time between the intact and the robotic knees, respectively,

$$\Delta P_k = P_k^p - P_k^i,\ \Delta D_k = D_k^p - D_k^i. \tag{4}$$

We then define robot state $x_k \in \mathbb{R}^4$ to include the following four variables, the first two of which are shared between the human and the robot, $x_k = [\Delta\tau_k, \Delta\lambda_k, \Delta P_k, \Delta D_k]^T$.

As such, human states are available to the robot, but a subset of the robot state (the kinematic features) is not available to the human. This reasonably reflects practical situations in real life applications.

The robot control policy is a state feedback law from robot state to robot impedance control parameters, namely, control u_k consists of increments to the impedance control parameters (Eq. 2),

$$u_k = \left[\Delta K_k, \Delta B_k, (\Delta\theta_e)_k\right]^T. \tag{5}$$

Determining human control policy which originates from human neurocontrol circuit is a daunting task, especially now that sensing, perception and feedback control circuits are interrupted after a lower limb amputation. While agents' policy solutions are solved from Bellman optimality-like equations, a human cannot interpret and implement an MARL solution via their complicated neural circuits. But this is not our intention. We instead use the human control variable as an estimated input to the robot controller design, i.e., the robot is informed by an estimated human influence when they share the same goal of symmetrical walking.

As the very first step to demonstrate this idea, we consider the human user of the wearable robot intentionally or voluntarily walk at a reference step length, which may also be from an instructional feedback. For real life scenarios, such behavioral cues can change over time, task, or environment. But it is important to validate such formulation of considering human influence in the robot control design. Toward this end, we consider a practical reference cue denoted as a desired step length λ_o. We let v_k represent an endogenous control signal of the human which is a function of human physical and mental states that involve activities ranging from neural level to muscular and joint level. Accordingly, we represent the endogenous human influence solved from the MARL design as a step length λ_k^d. We therefore define human control as,

$$v_k = \lambda_k^d - \lambda_o. \tag{6}$$

Including this estimated human control into the problem of robot control design, we take into account human influence on human-robot walking performance while they share the same symmetrical walking goal.

Symmetrical walking as shared task goal. We consider the stage cost to be shared between the human and the robot.

$$U(x_k, v_k, u_k) = x_k^T R_x x_k + R_v v_k^2 + u_k^T R_u u_k + \mu h_k^2, \tag{7}$$

where $R_x \in \mathbb{R}^{4\times4}$ and $R_u \in \mathbb{R}^{3\times3}$ are positive semi-definite weighting matrices, R_v and μ are positive weighting constants, and $h_k = v_k - \bar{v}_k$. In the above, the difference between actual human step length λ_k^i and reference λ_o, denoted as $\bar{v}_k = \lambda_k^i - \lambda_o$, is considered practical and available.

In this formulation, the shared control objective is represented in two ways. First, the robot kinematic state variables in x_k are to match the intact knee, i.e., the robotic joint angle profile is to match that of the intact joint. Additionally, a walking symmetry measure is directly considered by the differences in step length and stance time between human and robot, commonly used gait symmetry measures [63, 64, 65]. The human is assumed to perceive the gait symmetry but do not have access to or understand robot kinematic data (peak knee deflection error ΔP_k and peak knee time error

ΔD_k). The works [25, 50] show from test subject data that human control and adaptation has a direct influence on robot kinematics through dynamic interactions [61]. To account for this dynamic learning phenomenon, this framework penalizes the estimation error $h_k = v_k - \bar{v}_k = \lambda_k^d - \lambda_k^i$ between human perceived difference in step length. An unexpected or undesired human action will increase the cost represented by h_k, while a diminishing h_k indicates that human input that originated from the neuromuscular system matches actual human performance. Minimal controller energy expenditures are also included in the cost structure.

The human-robot control objective to be solve by MARL as a function of the state variables and the the controls is formulated as an infinite horizon, discounted cost $Q(x_k, v_k, u_k) = \sum_{j=k}^{\infty} \gamma^{j-k} U(x_j, v_j, u_j)$ where γ ($0 < \gamma < 1$) is the discount factor for the infinite-horizon problem.

Solutions to control actions. Let $v_k = \pi_v(z_k)$ and $u_k = \pi_u(x_k)$, where π_v and π_u are the human and the robot control policies solved from cMARL, respectively. Solving optimal control policies π_v^* and π_u^* requires solving the optimal Q function that satisfies Bellman optimality equation:

$$Q^*(x_k, v_k, u_k) = U(x_k, v_k, u_k) + \gamma Q^*(x_{k+1}, \pi_v^*(z_{k+1}), \pi_u^*(x_{k+1})). \tag{8}$$

When using a neural network-based actor-critic solution framework, we approximately solve the optimal control problem using the following iterative procedure (i is the iteration index),

$$Q_{i+1}(x_k, v_k, u_k) = U(x_k, v_k, u_k) + \gamma Q_i(x_{k+1}, \pi_{v_i}(z_{k+1}), \pi_{u_i}(x_{k+1})), \tag{9}$$

where $\pi_{v_i}(z_k) = arg\min_{v_k} Q_i(x_k, v_k, u_k), \pi_{u_i}(x_k) = arg\min_{u_k} Q_i(x_k, v_k, u_k)$, and $Q_i(x_k, v_k, u_k)$ are iterative actor policies and iterative Q value function.

During training, the actor and critic back-propagate their respective squared error to update their weights. The prediction error of actor $e_{a_v,k}, e_{a_u,k} \in \mathbb{R}$ is,

$$e_{a_v,k} = e_{a_u,k} = \frac{1}{2}(Q_i\left(x_k, v_k, u_k\right))^2. \tag{10}$$

The prediction error for the critic $e_{c,k}$ is formulated based on the Bellman error,

$$\epsilon_{c,k} = U + \gamma Q_i(x_{k+1}, \pi_{v_i}(z_{k+1}), \pi_{u_i}(x_{k+1})) - Q_{i+1}\left(x_k, v_k, u_k\right), \tag{11}$$

and the critic neural network is trained to minimize $e_{c,k} = \frac{1}{2}\epsilon_{c,k}^2$.

The optimal state-action cost-to-go function $Q^*(x_k, v_k, u_k)$ is approximated by a neural network via backpropagation to minimize the Bellman error on the shared cost signal, not on a local cost signal for either the human or the robot as the human and the robot are strongly coupled. The human control and the robot control, in turn, are learned via backpropagation from the shared cost signal, which has been shown successful in learning several complex, real-time control solutions. Further details on the cMARL automatic control solution is provided in Appendix B.3.

5 Experiment

We conduct design evaluations of the proposed cMARL for symmetrical walking using OpenSim, a well-established open source biomechanical modeling tool for conducting biomechanics research and motor control science [66]. The robot knee control is realized within an FS-IC framework. In real life, initial impedance parameters can be selected based on manufacturer and/or rehabilitation clinician's recommendations. Similar care is given to OpenSim simulated walking yet in all evaluations, the initial impedance values vary from a large range of settings for fair examination. As in [46], we enforce realistic safety constraints to prevent the human from stumbling or falling. Impedance parameters are reset to initial impedance values if any of the state variables exceed the safety bounds. The safety protocols followed in this work are fully described in Appendix A.4.

To make the simulations reflective of real world conditions, sensor and actuator noise data extracted from real human experimental testing sessions is applied to all the simulations in this study. Appendix A.2 provides the complete procedure followed of extracting noise data from experiments involving human subjects and injecting it into all the simulations.

In this section, we provide results of a large set of simulation studies aiming at answering the following questions: 1) Does our cMARL solution framework provide better performance than state-of-the-art baselines, including MADDPG and COMA? 2) Does including human influence in robot

control design accelerate learning and improve success rate of policy? 3) Is our cMARL applicable to different and realistic walking tasks? To provide answers, we show three sets of evaluations: benchmark, ablation and reliability. Benchmark and ablation evaluations are based on a level ground walking task with a pace of 1m/s. Reliability evaluations are based two new walking tasks: slope walking on a 11.5 degree ramp and level ground walking at an increased pace of 1.12m/s.

A trial consists of 300 gait cycles of continuous walking, which is chosen according to published results based on single-agent design studies using simulations and human experiments. Each gait cycle will generate one data tuple $(x_k, u_k, v_k, U_v, U_u, x_{k+1})$. A trial is considered successful if all state variables in x_k reach their respective error tolerance bound (Appendix A.4, Table 1, bottom row). Additionally, one gait phase at a time, if 8 out of 10 consecutive gaits meet the state error tolerance bound condition, training has converged. If all 4 phases have converged within 300 gait cycles, the trial is a success. Results reported in this paper are based on 16 trials of OpenSim simulations for each study below.

Different code implementation of different RL algorithms is known to cause large variances and impact on performance. To minimize such potential influence, all algorithms are implemented using Matlab [67], and all actor-critic implementations in all comparisons share the same code base [46, 68].

Benchmark study. MADDPG and COMA differ from our proposed cMARL approach toward symmetry walking (Figure 1, Right). MARL problems can vary greatly, so are expected for their solutions [69, 70].

To perform benchmark studies, we tailor MADDPG and COMA, respectively as follows. Refer to Figure 7. For MADDPG, we allow human and robot have different stage cost as $U_v = z^T R_z z + R_v v^2 + \mu h^2$ and $U_u = (x^-)^T (R_{x^-}) x^- + u^T R_u u$, respectively where x^- only includes the robotic joint kinematic variables. These stage costs respectively go to update critic Q_v and Q_u for human and robot. Human policy (π_v) (function of z) and robot policy (π_u) (function of x^-) are determined accordingly. For COMA, we train a single centralized critic with joint stage cost $U = U_v + U_u$ where U_v, U_u are the same as in MADDPG. Policies will execute locally. We update the policies using respective TD errors in place of a non-biased advantage function [57] because we are dealing with a strongly coupled human-robot system. Therefore it does not make sense to implement the counterfactual advantage function as in COMA. Also note that, we use real-time training instead of batch memory because collecting batch data for policy update requires an amputee subject to endure multiple gait cycles before one learning update. This may not be realistic for practical use. Nonetheless, an evaluation is performed (Appendix D). Details on the derivations and implementations of the tailored MADDPG and COMA are in Appendix C.

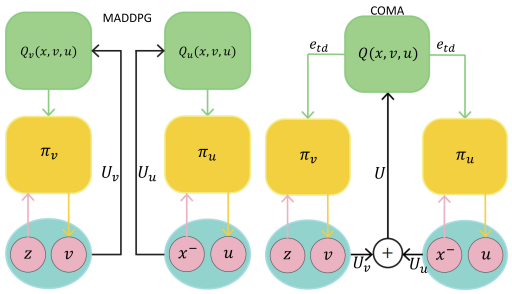

Figure 2: MADDPG (left) and COMA (right) tailored to solve the human-robot symmetrical walking problem.

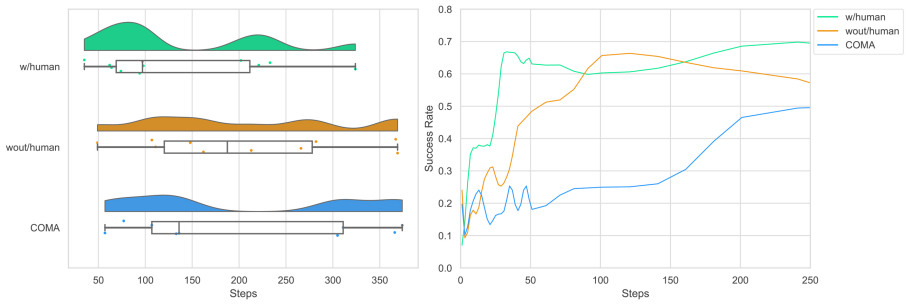

Figure 3: The total number of steps needed to reach convergence in training (left). The success rate during evaluation (right).

Each of the three approaches used the same random seeds during training and evaluation. Evaluation trials are performed on 5 randomly selected ones out of the 16 training trials, and evaluations are conducted every 2 gaits during the first 50 gaits and then every 10 gaits afterwards, each using the latest policy at the time of evaluation. Due to very low success rate of the tailored MADDPG method and very low chance of retrieving a successful policy (Figure 4, grey lines), also that its current implementation may be viewed as a decomposed COMA, we have removed MADDPG from evaluation.

Figure 3 shows that our cMARL solution has the best convergence profile and success rate over the baselines. Training and evaluation results are shown in Figures 4 and 5 which compare the environment sample efficiency as well as algorithm performances. To answer the first question based on benchmarking, our cMARL solution outperforms the baselines both in terms of kinematic and symmetry measurements as shown in Figures 4 and 5, center and right panels for training and evaluation, respectively. Additionally, the left panels of Figures 3, 4 and 5 show that our cMARL solution outperforms benchmarks with at least 30% less environment samples.

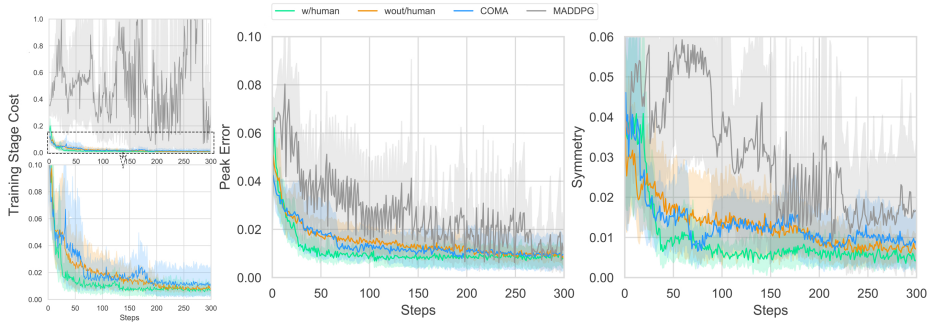

Figure 4: Learning curves of stage cost during training (left), peak angle error (middle) and symmetry of step length (right) for benchmark (w/human, COMA, MADDPG) and ablation (w/human, wout/human) studies. Each learning curve is averaged over 16 different random seeds and shaded by their respective 95% confidence interval.

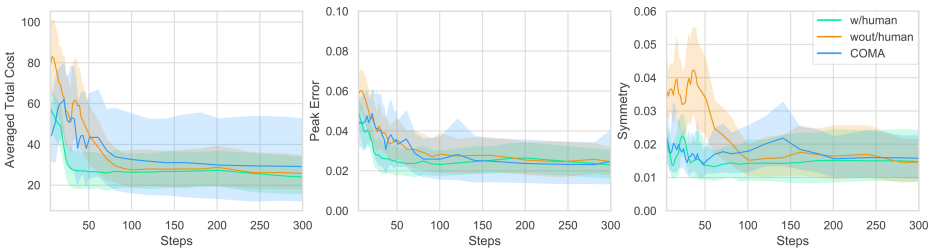

Figure 5: Learning curves of averaged total cost (terms associated with human influence removed) during evaluation (left), peak angle error (middle) and symmetry of step length (right) for benchmark (w/human, COMA) and ablation (w/human, wout/human) studies. Each (smoothed) learning curve is averaged over 5 different random seeds and shaded by their respective 95% confidence interval.

Ablation study. To gain insights on how human control influences human-robot walking performance, an ablation study is carried out with direct human influenced terms removed, including h_k and v_k, from the problem formulation. This only leaves the kinematics and symmetry measurements in the stage cost. Similar to the benchmark study, the ablation study has a training session and evaluation session. For comparable results, the total cost in evaluation is obtained as $U = (x)^T (R_x) x + u^T R_u u$. Figures 4 and 5 show that with an estimated human control influence accounted for in the robot control design, our cMARL solution (the green curves) outperforms the one without it (orange curves). Treating human and robot as collaborating agents toward a shared performance goal, our cMARL solution approach achieves increased success rate and accelerated learning speed (Figure 3). This result makes sense as an estimated human control provides a predictive signal to the robot control which aims at duplicating the intact human joint movement. Additional information on the quality of estimated human control is given in Appendix E.

Reliability study. Driven by making the proposed cMARL method practical and useful in real life, we setup two different walking tasks: (1) slope walking (11.5 degree ramp) and (2) walking at an increased pace (1.12m/s). They will result in different walking patterns from those used in the baseline study, and thus different knee joint profiles. For slope walking, knee flexion will be more pronounced during the stance phase since it walks inclined. In the case of faster pace, stance time will be compressed. To carry out the tests, the same training procedure is used as in benchmark and ablation studies.

Figure. 6 shows the performance of cMARL during slope walking and increased pace walking tasks. Performance of the two new tasks follow the same trend as that of the level ground walking at a nominal pace. These results again validate our design approach of using the intact knee movement trajectory as the target for the robotic knee to copy. By doing so, we have removed a major control design barrier in the way of performing different walking tasks by automatic control.

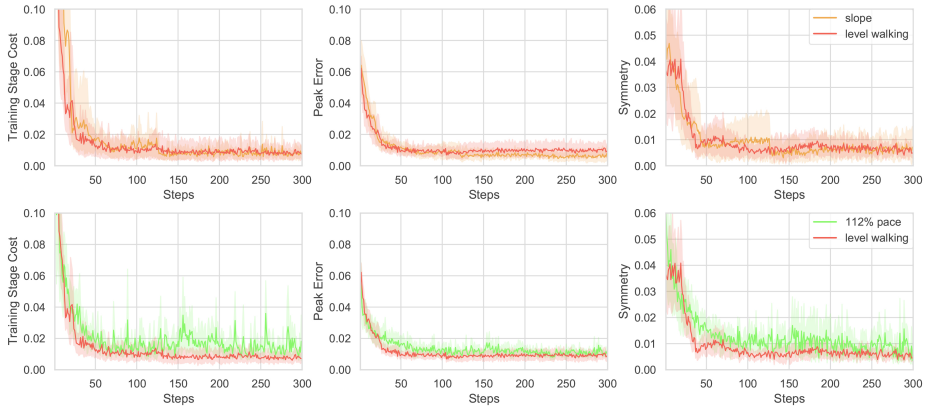

Figure 6: Learning curves of stage cost during training (Left), peak angle error (Middle) and symmetry of step length (Right) for different walking tasks (level ground walking at increased pace and ramp walking). Each learning curve is averaged over 16 different random seeds and shaded by their respective 95% confidence interval.

6 Conclusion and discussion

1) In this work, we develop a novel cMARL framework towards systematically integrating the human and robot as collaborative agents to achieve normative walking in real world scenarios. With reaching symmetric locomotion as shared control performance goal, we demonstrate expedited learning time and improved symmetry of walking. Although this new collaborative control framework can in principle be expanded to considering additional locomotion performance measures (which represents a fundamental and unique challenge in robotic lower limb prosthesis), additional work is needed to fully and appropriately represent endogenous human controls and their effect on human-robot system during walking. Further extensive simulation studies and human subject testings are also needed to make the proposed control framework fully utilizable in real life situations. 2) Symmetry is selected as the shared goal for the collaborating agents because asymmetric walking has been linked to secondary health complications including back pain and osteoarthiritis [71, 72, 73]. Although human-robot walking performance goals are difficult to systematically catalogue, additional considerations such as human energy expenditure and embodiment of the robot into the human will be considered in future works. 3) By breaking apart the shared cost for the human and the robot (cf. Section 5.1), the symmetrical walking task is treated and evaluated by MADDPG and COMA, respectively. Simulation results show that the factorization-based CTDE paradigm struggles to address the human-robot problem. The observed performance issues with factorization likely stem from the intrinsic coupling between the human and robot agents. 4) While ensuring human user safety has been carefully considered during control design, additional analysis of important properties such as convergence of learning, (sub)optimality of control policy, and human-robot closed-loop stability is still needed for this framework. Encouragingly, previous related works [74, 75, 76] indicate that these theoretical results are very likely to be provable.

References

[1] Markus Kuderer, Shilpa Gulati, and Wolfram Burgard. Learning driving styles for autonomous vehicles from demonstration. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2641–2646. IEEE, 2015.

[2] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[3] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[5] Martin Lawitzky, Alexander Mörtl, and Sandra Hirche. Load sharing in human-robot cooperative manipulation. In *19th International Symposium in Robot and Human Interactive Communication*, pages 185–191. IEEE, 2010.

[6] Hongbo Wang and Kazuhiro Kosuge. Control of a robot dancer for enhancing haptic human-robot interaction in waltz. *IEEE transactions on haptics*, 5(3):264–273, 2012.

[7] Antoine Bussy, Pierre Gergondet, Abderrahmane Kheddar, François Keith, and André Crosnier. Proactive behavior of a humanoid robot in a haptic transportation task with a human partner. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 962–967. IEEE, 2012.

[8] David Vogt, Simon Stepputtis, Steve Grehl, Bernhard Jung, and Heni Ben Amor. A system for learning continuous human-robot interactions from human-human demonstrations. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2882–2889. IEEE, 2017.

[9] Shuhei Ikemoto, Takashi Minato, and Hiroshi Ishiguro. Analysis of physical human–robot interaction for motor learning with physical help. *Applied Bionics and Biomechanics*, 5(4):213–223, 2008.

[10] Biao Ma, Yulong Liu, Xiaoxiang Na, Yahui Liu, and Yiyong Yang. A shared steering controller design based on steer-by-wire system considering human-machine goal consistency. *Journal of the Franklin Institute*, 356(8):4397–4419, 2019.

[11] Seyed Farokh Atashzar, Mahya Shahbazi, and Rajni V Patel. Haptics-enabled interactive neurorehabilitation mechatronics: classification, functionality, challenges and ongoing research. *Mechatronics*, 57:1–19, 2019.

[12] S Farokh Atashzar, Hsien-Yung Huang, Fulvia Del Duca, Etienne Burdet, and Dario Farina. Energetic passivity decoding of human hip joint for physical human-robot interaction. *IEEE Robotics and Automation Letters*, 5(4):5953–5960, 2020.

[13] Tingfang Yan, Marco Cempini, Calogero Maria Oddo, and Nicola Vitiello. Review of assistive strategies in powered lower-limb orthoses and exoskeletons. *Robotics and Autonomous Systems*, 64:120–136, 2015.

[14] Shraddha Srivastava, Pei-Chun Kao, Seok Hun Kim, Paul Stegall, Damiano Zanotto, Jill S Higginson, Sunil K Agrawal, and John P Scholz. Assist-as-needed robot-aided gait training improves walking function in individuals following stroke. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(6):956–963, 2014.

[15] Andres Martinez, Brian Lawson, Christina Durrough, and Michael Goldfarb. A velocity-field-based controller for assisting leg movement during walking with a bilateral hip and knee lower limb exoskeleton. *IEEE Transactions on Robotics*, 35(2):307–316, 2018.

[16] Dylan P Losey, Craig G McDonald, Edoardo Battaglia, and Marcia K O'Malley. A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction. *Applied Mechanics Reviews*, 70(1), 2018.

[17] Thomas B Sheridan. Human–robot interaction: status and challenges. *Human factors*, 58(4):525–532, 2016.

[18] Stephanie Huang, Jeffrey P Wensman, and Daniel P Ferris. Locomotor adaptation by transtibial amputees walking with an experimental powered prosthesis under continuous myoelectric control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(5):573–581, 2015.

[19] Charles T Leonard, RL Craik, and CA Oatis. The neurophysiology of human locomotion. In *Gait Analysis: Theory and Application*. Mosby-Year, 1995.

[20] Gentaro Taga, Yoko Yamaguchi, and Hiroshi Shimizu. Self-organized control of bipedal locomotion by neural oscillators in unpredictable environment. *Biological cybernetics*, 65(3):147–159, 1991.

[21] Ryan T Roemmich and Amy J Bastian. Closing the loop: from motor neuroscience to neurorehabilitation. *Annual review of neuroscience*, 41:415–429, 2018.

[22] Agostino De Santis, Bruno Siciliano, Alessandro De Luca, and Antonio Bicchi. An atlas of physical human–robot interaction. *Mechanism and Machine Theory*, 43(3):253–270, 2008.

[23] Panagiota Tsarouchi, Sotiris Makris, and George Chryssolouris. Human–robot interaction review and challenges on task planning and programming. *International Journal of Computer Integrated Manufacturing*, 29(8):916–931, 2016.

[24] Minhan Li, Yue Wen, Xiang Gao, Jennie Si, and He Helen Huang. Towards expedited impedance tuning of a robotic prosthesis for personalized gait assistance by reinforcement learning control. *arXiv preprint arXiv:2006.06518*, 2020.

[25] Yue Wen, Minhan Li, Jennie Si, and He Huang. Wearer-prosthesis interaction for symmetrical gait: a study enabled by reinforcement learning prosthesis control. *IEEE transactions on neural systems and rehabilitation engineering*, 28(4):904–913, 2020.

[26] Kenton R Kaufman, Serena Frittoli, and Carlo A Frigo. Gait asymmetry of transfemoral amputees using mechanical and microprocessor-controlled prosthetic knees. *Clinical Biomechanics*, 27(5):460–465, 2012.

[27] Rui Huang, Zhinan Peng, Hong Cheng, Jiangping Hu, Jing Qiu, Chaobin Zou, and Qiming Chen. Learning-based walking assistance control strategy for a lower limb exoskeleton with hemiplegia patients. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2280–2285. IEEE, 2018.

[28] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[30] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[31] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, 2018.

[32] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[33] Seungmoon Song, Łukasz Kidziński, Xue Bin Peng, Carmichael Ong, Jennifer Hicks, Sergey Levine, Christopher G Atkeson, and Scott L Delp. Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation. *Journal of neuroengineering and rehabilitation*, 18(1):1–17, 2021.

[34] Qingsong Ai, Zemin Liu, Wei Meng, Quan Liu, and Sheng Q Xie. Machine learning in robot assisted upper limb rehabilitation: A focused review. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[35] Weidong Wang, Chengjin Du, Wei Wang, and Zhijiang Du. A pso-optimized fuzzy reinforcement learning method for making the minimally invasive surgical arm cleverer. *IEEE Access*, 7:48655–48670, 2019.

[36] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[37] A Isidori. *Nonlinear Control Systems: An Introduction*. Springer Verlag, Berlin, Germany, 1985.

[38] Russell Enns and Jennie Si. Apache helicopter stabilization using neural dynamic programming. *Journal of guidance, control, and dynamics*, 25(1):19–25, 2002.

[39] Russell Enns and Jennie Si. Helicopter trimming and tracking control using direct neural dynamic programming. *IEEE transactions on neural networks*, 14(4):929–939, 2003.

[40] Russell Enns and Jennie Si. Helicopter trimming and tracking control using direct neural dynamic programming. *IEEE transactions on neural networks*, 14(4):929–939, 2003.

[41] Chao Lu, Jennie Si, and Xiaorong Xie. Direct heuristic dynamic programming for damping oscillations in a large power system. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(4):1008–1013, 2008.

[42] Wentao Guo, Feng Liu, Jennie Si, Dawei He, Ronald Harley, and Shengwei Mei. Approximate dynamic programming based supplementary reactive power control for dfig wind farm to enhance power system stability. *Neurocomputing*, 170:417–427, 2015.

[43] Jingyi Zhang, Yonghong Luo, Boya Wang, Chao Lu, Jennie Si, and Jie Song. Deep reinforcement learning for load shedding against short-term voltage instability in large power systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[44] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *arXiv preprint arXiv:1707.01495*, 2017.

[45] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[46] Yue Wen, Jennie Si, Xiang Gao, Stephanie Huang, and He Helen Huang. A new powered lower limb prosthesis control framework based on adaptive dynamic programming. *IEEE transactions on neural networks and learning systems*, 28(9):2215–2220, 2016.

[47] Yue Wen, Jennie Si, Andrea Brandt, Xiang Gao, and He Huang. Online reinforcement learning control for the personalization of a robotic knee prosthesis. *IEEE transactions on cybernetics*, 2019.

[48] Wentao Liu, Junmin Zhong, Ruofan Wu, Bretta L Fylstra, Jennie Si, and He Helen Huang. Inferring human-robot performance objectives during locomotion using inverse reinforcement learning and inverse optimal control. *IEEE Robotics and Automation Letters*, 2022.

[49] Yue Wen, Andrea Brandt, Ming Liu, He Huang, and Jennie Si. Comparing parallel and sequential control parameter tuning for a powered knee prosthesis. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1716–1721. IEEE, 2017.

[50] Yue Wen, Andrea Brandt, Jennie Si, and He Helen Huang. Automatically customizing a powered knee prosthesis with human in the loop using adaptive dynamic programming. In *2017 International Symposium on Wearable Robotics and Rehabilitation (WeRob)*, pages 1–2. IEEE, 2017.

[51] Yue Wen, Xiang Gao, Jennie Si, Andrea Brandt, Minhan Li, and He Helen Huang. Robotic knee prosthesis real-time control using reinforcement learning with human in the loop. In *International Conference on Cognitive Systems and Signal Processing*, pages 463–473. Springer, 2018.

[52] Xiang Gao, Jennie Si, Yue Wen, Minhan Li, and He Huang. Reinforcement learning control of robotic knee with human-in-the-loop by flexible policy iteration. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[53] Minhan Li, Yue Wen, Xiang Gao, Jennie Si, and He Huang. Toward expedited impedance tuning of a robotic prosthesis for personalized gait assistance by reinforcement learning control. *IEEE Transactions on Robotics*, 2021.

[54] Minhan Li, Xiang Gao, Yue Wen, Jennie Si, and He Helen Huang. Offline policy iteration based reinforcement learning controller for online robotic knee prosthesis parameter tuning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2831–2837. IEEE, 2019.

[55] Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

[56] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017.

[57] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[58] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

[59] Peng Peng, Ying Wen, Yaodong Yang, Quan Yuan, Zhenkun Tang, Haitao Long, and Jun Wang. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*, 2017.

[60] Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. Stabilising experience replay for deep multi-agent reinforcement learning. In *International conference on machine learning*, pages 1146–1155. PMLR, 2017.

[61] Neville Hogan. Impedance control: An approach to manipulation: Part i—theory. 1985.

[62] Frank Sup, Amit Bohara, and Michael Goldfarb. Design and control of a powered transfemoral prosthesis. *The International journal of robotics research*, 27(2):263–273, 2008.

[63] Slavka Viteckova, Patrik Kutilek, Zdenek Svoboda, Radim Krupicka, Jan Kauler, and Zoltan Szabo. Gait symmetry measures: A review of current and prospective methods. *Biomedical Signal Processing and Control*, 42:89–100, 2018.

[64] Antonio Nardone, Marco Godi, Margherita Grasso, Simone Guglielmetti, and Marco Schieppati. Stabilometry is a predictor of gait performance in chronic hemiparetic stroke patients. *Gait & Posture*, 30(1):5–10, 2009.

[65] Chris Mizelle, Mary Rodgers, and Larry Forrester. Bilateral foot center of pressure measures predict hemiparetic gait velocity. *Gait & Posture*, 24(3):356–363, 2006.

[66] Scott L Delp, Frank C Anderson, Allison S Arnold, Peter Loan, Ayman Habib, Chand T John, Eran Guendelman, and Darryl G Thelen. Opensim: open-source software to create and analyze dynamic simulations of movement. *IEEE transactions on biomedical engineering*, 54(11):1940–1950, 2007.

[67] MATLAB. *version 7.10.0 (R2022a)*. The MathWorks Inc., Natick, Massachusetts, 2022.

[68] Xiang Gao, Jennie Si, Yue Wen, Minhan Li, and He Helen Huang. Knowledge-guided reinforcement learning control for robotic lower limb prosthesis. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 754–760. IEEE, 2020.

[69] Afshin OroojlooyJadid and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

[70] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

[71] M Jason Highsmith, Casey R Andrews, Claire Millman, Ashley Fuller, Jason T Kahle, Tyler D Klenow, Katherine L Lewis, Rachel C Bradley, and John J Orriola. Gait training interventions for lower extremity amputees: a systematic literature review. *Technology & Innovation*, 18(2-3):99–113, 2016.

[72] Alberto Esquenazi. Gait analysis in lower-limb amputation and prosthetic rehabilitation. *Physical Medicine and Rehabilitation Clinics*, 25(1):153–167, 2014.

[73] George NS Marinakis. Interlimb symmetry of traumatic unilateral transtibial amputees wearing two different prosthetic feet in the early rehabilitation stage. *Journal of rehabilitation research and development*, 41(4):581–590, 2004.

[74] Feng Liu, Jian Sun, Jennie Si, Wentao Guo, and Shengwei Mei. A boundedness result for the direct heuristic dynamic programming. *Neural Networks*, 32:229–235, 2012.

[75] Yury Sokolov, Robert Kozma, Ludmilla D Werbos, and Paul J Werbos. Complete stability analysis of a heuristic approximate dynamic programming control design. *Automatica*, 59:9–18, 2015.

[76] Xiang Gao. *Data-Efficient Reinforcement Learning Control of Robotic Lower-Limb Prosthesis With Human in the Loop*. PhD thesis, Arizona State University, 2020.

[77] Daniel Jacobs. From the ground up: Building a passive dynamic walker model, 2014.

[78] Neville Hogan. Impedance control: An approach to manipulation: Part ii—implementation. 1985.

[79] Neville Hogan. Impedance control: An approach to manipulation: Part iii—applications. *Journal of Dynamic Systems Measurement and Control-transactions of The Asme*, 107:17–24, 1985.

[80] Ming Liu, Fan Zhang, Philip Datseris, and He Helen Huang. Improving finite state impedance control of active-transfemoral prosthesis using dempster-shafer based state transition rules. *Journal of Intelligent & Robotic Systems*, 76(3-4):461–474, 2014.

[81] Michael Windrich, Martin Grimmer, Oliver Christ, Stephan Rinderknecht, and Philipp Beckerle. Active lower limb prosthetics: a systematic review of design issues and solutions. *Biomedical engineering online*, 15(3):5–19, 2016.

[82] David J Sanderson and Philip E Martin. Lower extremity kinematic and kinetic adaptations in unilateral below-knee amputees during walking. *Gait & posture*, 6(2):126–136, 1997.

[83] Roberto E Quesada, Joshua M Caputo, and Steven H Collins. Increasing ankle push-off work with a powered prosthesis does not necessarily reduce metabolic rate for transtibial amputees. *Journal of Biomechanics*, 49(14):3452–3459, 2016.

[84] Philippe Malcolm, Roberto E Quesada, Joshua M Caputo, and Steven H Collins. The influence of push-off timing in a robotic ankle-foot prosthesis on the energetics and mechanics of walking. *Journal of neuroengineering and rehabilitation*, 12(1):1–15, 2015.

[85] Yue Wen, Minhan Li, Jennie Si, and He Helen Huang. Wearer-prosthesis interaction for symmetrical gait: A study enabled by reinforcement learning prosthesis control. *IEEE transactions on Neural Systems and Rehabilitation*, 2020.

[86] Jennie Si, Andrew G Barto, Warren B Powell, and Don Wunsch. *Handbook of learning and approximate dynamic programming*, volume 2. John Wiley & Sons, 2004.

[87] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

[88] Clautilde Nguiadem, Maxime Raison, and Sofiane Achiche. Motion planning of upper-limb exoskeleton robots: A review. *Applied Sciences*, 10(21):7626, 2020.

[89] Mingyu Cai, Mohammadhosein Hasanbeig, Shaoping Xiao, Alessandro Abate, and Zhen Kan. Modular deep reinforcement learning for continuous motion planning with temporal logic. *IEEE Robotics and Automation Letters*, 6(4):7973–7980, 2021.

[90] Fei Ye, Shen Zhang, Pin Wang, and Ching-Yao Chan. A survey of deep reinforcement learning algorithms for motion planning and control of autonomous vehicles. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1073–1080. IEEE, 2021.

[91] Ricardo B Grando, Junior C de Jesus, Victor A Kich, Alisson H Kolling, Nicolas P Bortoluzzi, Pedro M Pinheiro, Armando A Neto, and Paulo LJ Drews. Deep reinforcement learning for mapless navigation of a hybrid aerial underwater vehicle with medium transition. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1088–1094. IEEE, 2021.

[92] Yan Du, Helia Zandi, Olivera Kotevska, Kuldeep Kurte, Jeffery Munk, Kadir Amasyali, Evan Mckee, and Fangxing Li. Intelligent multi-zone residential hvac control strategy based on deep reinforcement learning. *Applied Energy*, 281:116117, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A] We don't have theoretical analysis

(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Detail instructions in supplemental material

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In supplement file

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] In supplement file

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] In supplement file

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]