Generating Realistic Arm Movements in Reinforcement Learning: A Quantitative Comparison of Reward Terms and Task Requirements

Jhon P.F. Charaja*[1], Isabell Wochner[2], Pierre Schumacher[1,3], Winfried Ilg[1], Martin Giese[1],
Christophe Maufroy[4,5], Andreas Bulling[6], Syn Schmitt[7], Georg Martius[3,8], and Daniel F.B. Haeufle[1,2]

***Abstract*— Mimicking of human-like arm movement characteristics involves considering three factors during control policy synthesis: (a) task requirements, (b) noise during movement execution, and (c) optimality principles. Previous studies showed that when these factors (a-c) are considered individually, it is possible to synthesize arm movements that either kinematically match experimental data or reproduce the stereotypical triphasic muscle activation pattern. However, no quantitative comparison has assessed the realism of arm movements generated by each factor, nor has it been determined whether combining these factors results in movements with human-like kinematic characteristics and the triphasic muscle pattern. To investigate this, we used reinforcement learning to learn a control policy for a musculoskeletal arm model, aiming to discern which combination of factors (a-c) results in realistic arm movements according to four frequently reported stereotypical characteristics. Our findings indicate that incorporating velocity and acceleration requirements into the reaching task, employing reward terms that minimize mechanical work, hand jerk, and control effort, along with the inclusion of noise during movement, leads to realistic human arm movements by reinforcement learning. We expect that the gained insights will help in the future to better predict desired arm movements and corrective forces in wearable assistive devices.**

I. Introduction

In aging societies, the number of people benefiting from motor rehabilitation is on the rise [1]. Assistive devices promise support in activities of daily living, e.g., reaching for tools and objects [2]. The design and control of assistive devices would benefit from models accurately predicting human movement. Reinforcement learning in combination with biomechanical models can lead to the emergence of natural characteristics, such as gait kinematics [3] and hand trajectories [4]. However, this requires identifying reward terms and task requirements that lead to realistic movements.

Arm-reaching movements exhibit highly stereotypical kinematics and temporal characteristics. Important characteristics documented in literature are: (i) roughly straight hand trajectories, (ii) bell-shaped tangential velocity profiles [5], [6], (iii) triphasic muscle activation pattern [7], [8], i.e., the alternating activation of agonist and antagonist muscles, and (iv) linear relationship between movement time (MT) and index of difficulty (ID) (a.k.a Fitts's law) [9]. Several optimality principles have been proposed for deterministic prediction of arm-reaching movements, such as minimal work, jerk or muscular effort [10], [11]. Flash et al. [11] found that minimization of hand jerk predicts characteristics (i&ii) in point-to-point movements. Wochner et al. [12] indicated that minimization of mechanical work, jerk, and muscle stimulation command (effort) predicts characteristics (i&ii) in point-to-manifold movements. Finally, Ueyama et al. [13] demonstrated that minimization of control effort and consideration of position, velocity, and force requirements in the reaching task predict characteristics (i-iii). As a stochastic approach, Fischer et al. [4] applied constant and signal-dependent noise of muscle stimulation amplitude. Combined with minimization of movement time they were able to reproduce characteristics (i&ii&iv) on point-to-point movements. To our knowledge, no simulation approach has investigated all four characteristics.

More precisely, three factors influence the resulting behavior of the control policy to generate human characteristics of arm movement: (a) the chosen task requirements, (b) inclusion of noise during movement execution and (c) the chosen optimality principles. Some of these factors have been evaluated based on their ability to generate kinematic characteristics that match experimental data, while others evaluated the emergence of the triphasic muscle activation pattern. However, no quantitative comparison has been conducted on the realism of the arm movement generated by each factor; as well as whether a partial or total combination of all factors results in arm movements with human-like kinematic and muscle activation pattern.

The purpose of this study is to investigate which combination of factors (a-c) result in realistic arm movements according to the four stereotypical characteristics (i-iv) defined above. We test this using reinforcement learning to learn a control policy for a musculoskeletal arm model and systemically investigate a combination of (a) the chosen task requirements, (b) inclusion of noise during movement execution and (c) the chosen optimality principles with the aim of methodically evaluating their contribution—for the first time—in one model. We expect that the gained insights will help in the future to better predict desired movements and corrective forces in assistive devices.

*corresponding author: jhon.charaja-casas@uni-tuebingen.de, [1]Hertie Institute for Clinical Brain Research, and Centre for Integrative Neuroscience, University of Tübingen, Germany. [2]Institute of Computer Engineering (ZITI), Heidelberg University, Germany. [3]Max Planck Institute for Intelligent Systems, Tübingen, Germany. [4]Institut for Industrial Manufacturing and Management (IFF), University of Stuttgart, Germany. [5]Fraunhofer Institute for Manufacturing Engineering and Automation (IPA), Germany.[6]Institute for Visualization and Interactive Systems, University of Stuttgart, Germany. [7]Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart, Germany. [8]Department of Computer Science, University of Tübingen, Germany.

This work was financed by the Baden-Württemberg Stiftung in the scope of the AUTONOMOUS ROBOTICS project *iAssistADL* granted to WI, MG, SS, and DH. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Jhon Charaja.

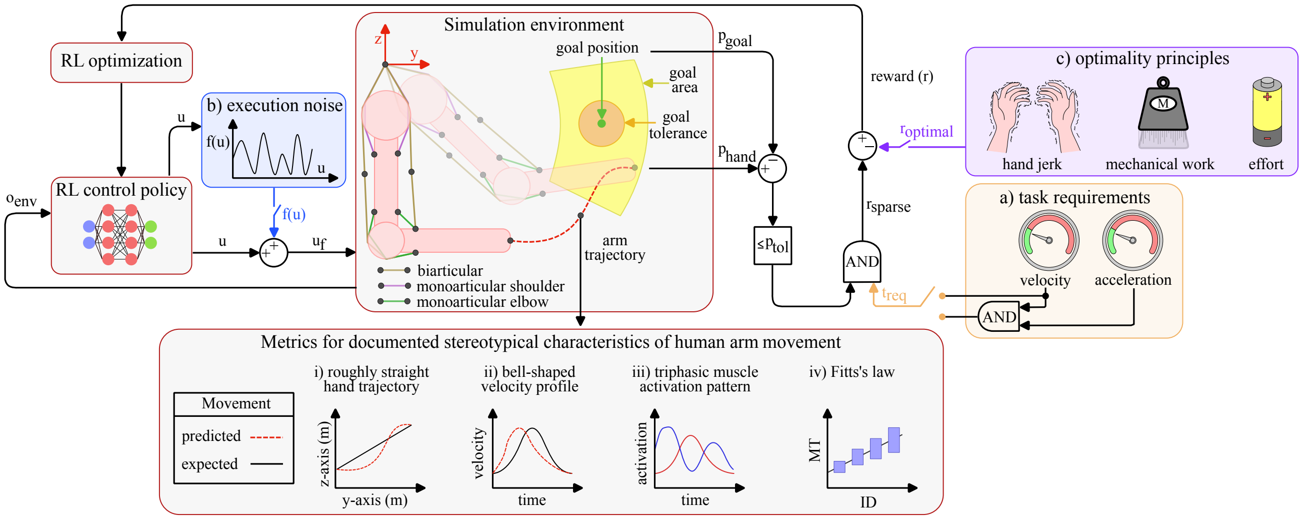

Fig. 1. Framework to systematically combine three factors that generate arm movements: (a) different task requirements, (b) inclusion of noise during movement execution, and (c) optimality principles grounded on the minimization of mechanical work, hand jerk and muscle stimulation command (effort). Each combination creates a unique learning environment with distinctive challenges and movement priorities: execution noise modifies the control commands, while optimality principles and additional task requirements shape the reward. Shown below are the metrics for documented stereotypical characteristics of human arm movement: (i) roughly straight hand trajectory, (ii) bell-shaped velocity profile, (iii) triphasic muscle activation pattern, and (iv) Fitts's law.

II. Methods

In a nutshell, the factors (a-c) are categorized into two primary domains: models and task requirements. We investigate four models: the *baseline* model only aims at minimizing movement time. In the other three models, the *baseline* model is combined with either *execution noise* (b), *optimality principles* (c) characterized by the minimization of mechanical work, hand jerk, and muscle stimulation commands, or a *hybrid* model that considers execution noise and optimality principles. For the task requirements (a) we consider three potential configurations: position only (pos), position and velocity (pos-vel), and position, velocity, and acceleration (pos-vel-acc), all aiming to fulfill respective kinematic constraints at the target location. Details will be given below. This organization facilitates the exploration of how different combinations of each factor (a-c) influence the behavior of the resulting control policy, as is illustrated in Figure 1. By finally analyzing the resulting movements according to the stereotypical characteristic (i-iv) of human arm movement, we can identify essential elements for generating arm movements that exhibit human characteristics without enforcing them explicitly as reward terms.

The simulation workflow requires: generation of muscle stimulation commands, simulation of human arm dynamics, calculating rewards, and using metrics for goal-oriented movements. In the following subsections, these will be described in detail.

A. Muscle stimulation commands

The RL agent utilizes Maximum a Posteriori Optimization (MPO) [14] combined with DEP-RL [15] for exploration; a novel approach that demonstrated robust performance controlling musculoskeletal systems. The MPO implementation follows the default settings provided by the TonicRL library [16][1], and the DEP-RL configuration mirrors the hyperparameters outlined for the same arm model in [15].

The RL agent undergoes training with the inclusion of execution noise (when activated) and random position targets sampled from an area determined by the arm model kinematics. The control policy (π) computes muscle stimulation commands (u) based on the current observation from the environment (o_{env}). The execution noise is introduced as $u_f = (1+\eta_1)u+\eta_2$, modifying the amplitude of control commands u, where η_1 represents the signal-dependent noise, η_2 represents the constant motor noise and u_f is the applied muscle stimulation command. Both noise signals are random Gaussian variables, each with a mean of 0. The standard deviation for η_1 is 0.103 and for η_2 is 0.185 [17].

B. Simulation of human arm dynamics

The physics engine MuJoCo [18] simulates the muscle activation dynamics resulting in the generation of muscle forces that drive the arm movements. In MuJoCo, a musculoskeletal model of the human arm with two degrees of freedom and six actuated muscles is available [18]. The original model was modified to generate arm movements in the sagittal plane (considering gravity). The position error is calculated between the tip of the forearm and the desired position (reaching goal). The RL environment considers the same initial conditions of the arm model for all episodes: $0°$ for the shoulder angle, $90°$ for the elbow angle, zero joint velocities and zero muscle activation level. The environment observation comprises Cartesian states[2], joint states[3], muscle states[4], mechanical work, hand jerk and the goal position. The agent's policy network generates control commands every 10 ms while the MuJoCo physics engine updates the

[1] except for the following parameters: batch size with 256, batch iteration with 30, steps before batches with $1e6$ and steps between batches with 50.

arm model every 2 ms and uses the same control command for five consecutive time steps.

C. Reward formulation

Previous studies have successfully generated reaching arm movements utilizing an optimal control framework [13], [12]. This methodology incorporates a terminal cost that penalizes deviations from a desired final state and an accumulated cost associated with the states and control commands during the trajectory [19]. Building upon this, the reward function consists of a sparse reward linked to the fulfillment of kinematic requirements at the end of the trajectory and a dense reward associated with executing optimal movements.

The immediate reward function is a combination of

$$r = c_1 r_{\text{sparse}} - c_2 r_{\text{optimal}}, \tag{1}$$

where $c_{1,2}$ represent weighting coefficients to establish priority during the generation of arm movements, r_{sparse} penalizes movement duration and r_{optimal} encourages optimal behavior based on the optimality principles described above. We select: $c_1 = 0.2$ and $c_2 = 0.8$.

Generally, arm-pointing movements are executed quickly. Consequently, previous research employed a constant negative reward for each step transition until the position requirement is met [4]. Additionally, Ueyama et al. [13] found that velocity and force requirements influence the stereotypical triphasic activation pattern. Considering these findings, we incorporate terminal velocity and acceleration (proportional to force in Cartesian space) requirements into the reaching task. Therefore, r_{sparse} depends on meeting the goal tolerance in position as well as the additional kinematic requirements.

$$r_{\text{sparse}} = \begin{cases} 0, & \text{if } \|p_{\text{hand}} - p_{\text{goal}}\| \leq p_{\text{tol}} \;\; \& \;\; \text{t}_{\text{req}} = \text{true} \\ -1, & \text{otherwise,} \end{cases} \tag{2}$$

where p_{hand} is the Cartesian hand position, p_{goal} is the Cartesian desired position, p_{tol} represents goal tolerance and t_{req} is the state of the additional task requirements as

$$\text{t}_{\text{req}} = \begin{cases} \text{true}, & \text{pos} \\ \|v\| \leq v_{\text{tol}}, & \text{pos-vel} \\ \|v\| \leq v_{\text{tol}} \;\; \& \;\; \|a\| \leq a_{\text{tol}}, & \text{pos-vel-acc} \end{cases}$$

where v, a are hand velocity and acceleration, $v_{\text{tol}}, a_{\text{tol}}$ are tolerance for velocity and acceleration. We choose v_{tol}, a_{tol} to be 10% of the maximum values observed solely under position task requirement: $v_{\text{tol}} = 20\,\frac{\text{cm}}{\text{s}}$ and $a_{\text{tol}} = 100\,\frac{\text{cm}}{\text{s}^2}$.

Furthermore, we consider four values for p_{tol} to address various difficulty levels in the reaching task. The difficulty associated with reaching movements can be calculated using the index of difficulty (ID) [9], defined as $\log_2\left(\frac{D}{W} + 1\right)$, where D represents goal distance and $W = 2p_{\text{tol}}$ represents endpoint variability. The values are selected conveniently to ensure that the resulting difficulty indices (ID $=$ 2 to 5) are integers: $D = 63\,\text{cm}$ (used for evaluation) and $p_{\text{tol}} = 10.5\,\text{cm},\ 4.5\,\text{cm},\ 2.1\,\text{cm},\ 1.0161\,\text{cm}$. For each combination of model and task requirements, one RL agent is trained for each tolerance value, resulting in a total of 48 RL agents.

Exclusively focusing on minimizing movement time will generate bang-bang control solutions with asymmetric velocity profiles [20]. Wochner et al. [12] found that bell-shaped velocity profiles emerge in point-to-manifold tasks only when optimal behavior considers the minimization of mechanical work, hand jerk, and muscle stimulation commands (related to muscular effort). As both Berret et al. [10] and Wochner et al. [12] suggested that it is crucial to consider a combination of optimality principles to tackle the redundancy problem, we therefore, consider the suggested combination r_{optimal} of three optimality principles as:

$$r_{\text{optimal}} = \frac{c_3 r_{\text{effort}} + c_4 r_{\text{jerk}} + c_5 r_{\text{work}}}{c_3 + c_4 + c_5}, \tag{3}$$

where $c_{3,4,5}$ set priority between optimality principles, r_{effort} is computed as mean value of muscle stimulation commands, r_{jerk} is estimated by finite difference computation between the current and one previous acceleration values, instantaneous work (power) r_{work} is computed as $|\dot{\phi}_1 \tau_1| + |\dot{\phi}_2 \tau_2|$, where $\dot{\phi}_{1,2}$ represents the angular velocity of shoulder and elbow, and $\tau_{1,2}$ indicates the torque of shoulder and elbow. We normalize each optimality principle by its observed maximum value: $r_{\text{jerk}} = 1000\,\frac{\text{m}}{\text{s}^3}$ and $r_{\text{work}} = 100\,\text{J}$. In pre-tests, we found that smooth muscle profiles were only achieved if all the three terms are considered with the following coefficients: $c_3 = 1$, $c_4 = 8$, and $c_5 = 1$.

D. Metrics for goal-oriented movements

We evaluate each agent in its training environment. The position target for all agents is positioned $29.5\,\text{cm}$ to the left and $55.7\,\text{cm}$ upward relative to the tip of the forelimb. We run 1000 rollouts for each agent to capture their average behavior. Since each test episode has a different movement time (MT), we temporally normalize the recorded data to make trajectory characteristics comparable. We recognize outliers by examining the velocity profile. Any trajectory with an integral velocity beyond the interquartile range (between the 25th and 75th percentiles) is excluded from consideration. The mean computed over the remaining rollouts is employed for the analysis of the movement's characteristics.

The performance of all trained agents is quantified with four metrics associated with the stereotypical characteristic (i-iv) observed in goal-oriented movements:

i. **Straight line deviation (p_{line}):** This metric reports the R-squared between the straight line from initial point to target and the actual hand trajectory.

ii. **Bell-shaped velocity profile (v_{bell}):** We determine the onset and offset of the velocity profile by the threshold $v > 0.1 v_{\text{max}}$ of the peak velocity. A Gaussian is fitted between onset and offset of the velocity profile. The Gaussian strictly considers peak velocity as amplitude, and the `fit` function of MatLab computes the mean

[2] position, velocity and acceleration of the hand, i.e., tip of the forearm.

[3] position, velocity, acceleration and jerk of each arm joint

[4] muscle activity, muscle forces, muscle lengths and muscle velocities

and standard deviation of the Gaussian. This metric (v_{bell}) reports the R-squared to indicate how bell-shaped each velocity profile is.

iii. **Triphasic muscle pattern ($u_{\textbf{triphasic}}$):** This metric analyzes the muscle activation pattern of each agonist-antagonist muscle pair in the arm. The aim is to capture if an antagonistic pair changes operation mode, e.g., if in the beginning elbow flexor is actively flexing the elbow and then, elbow extensor activity rises and elbow flexor activity falls to decelerate the movement, this is considered a second phase. We quantify this by evaluating if muscle activation slopes exchange directions and by the threshold $\Delta > 0.25\,\Delta_{\text{max}}$ and $\Delta > 1.5\,\text{e}-3$ of difference between them. If this is the case, it is considered a new phase of muscle activation. The metric verifies if the reported triphasic pattern in the literature [13] occurs in a muscle pair, assigning a score of 1 if true and 0 otherwise.

iv. **Fitts's law (R_F)):** This metrics reports the correlation coefficient R_F to indicate how strong the linearity is.

III. Results

Overall, incorporating velocity and acceleration requirements into the reaching task (pos-vel-acc), employing reward terms that minimize mechanical work, hand jerk, and control effort, along with the inclusion of noise during movement, leads to the most realistic arm movement according to the four proposed metrics (i-iv). Furthermore, increasing index of difficulty, from ID = 2 to 5, yields more bell-shaped velocity profiles and the emergence of the third phase in the muscle activation pattern. These results are presented in Table I, which shows the performance of all agents for each proposed metric across all difficulty indices (ID = 2 to 5). Note that in Table I, R_F only displays one value, as this metric utilizes all difficulty indices to determine how strong the linear relationship (correlation) between movement time (MT) and index of difficulty (ID) is (Fitts's law). Also, the velocity profiles obtained with only position task requirement (pos), do not reach the lower threshold of 10% of the peak velocity; consequently, these velocity profiles are not considered for the v_{bell} metric (displayed as solid line "-").

A. Straight line deviation (p_{line})

The best performance in terms of straight line deviation p_{line} across the majority of difficulty indices (ID = 2 to 5) for baseline, execution noise and optimality principles models is linked to pos-vel-acc task requirement (Tab. I). Conversely, the best performances of the hybrid model are distributed across pos and pos-vel task requirements. All hand trajectories with difficulty index ID = 5 are illustrated in Figure 2. The figure illustrates the progressive straightening of hand trajectories as more kinematic requirements are incorporated into the main task. It is noteworthy that even the worst p_{line} values (0.91, 0.92, ...) still represent lines that we would consider roughly straight. Consequently, solely relying on the p_{line} metric makes it implausible to indicate which combination will yield the most realistic hand trajectory.

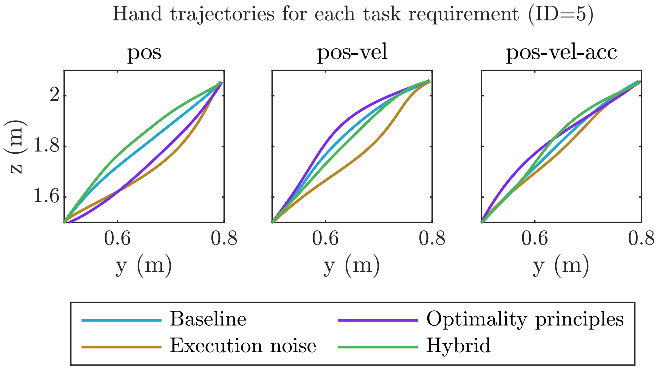

Fig. 2. Hand trajectories generated by all models, considering the three possible task requirements and difficulty index ID = 5.

B. Bell-shaped velocity profile (v_{bell})

The best performance in terms of bell-shaped velocity profile v_{bell} across the majority of difficulty indices (ID = 2 to 5) for baseline, execution noise, optimality principles and hybrid models is linked to pos-vel-acc task requirement (Tab. I). The hybrid model combined with pos-vel-acc task requirement, consistently exhibits the highest v_{bell} values, i.e., most bell-shaped velocity profiles, across all difficulty indices. In addition, Table I reveals a increasing trend of v_{bell} values with increasing index of difficulty. The velocity profile for ID = 5 of each model with pos-vel-acc task requirement are shown in Figure 3. The figure illustrates that all models align well with the right side of the Gaussian model, and fitting errors arise from the left side.

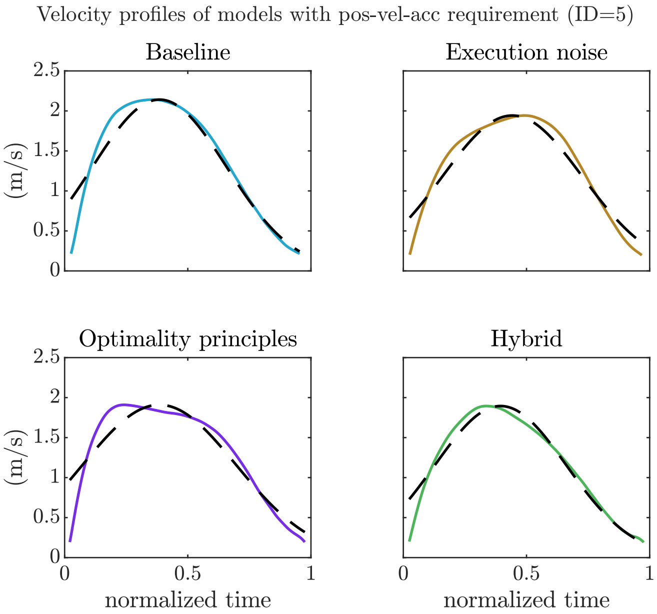

Fig. 3. Velocity profiles generated by each model, considering velocity and acceleration requirements into main task and difficulty index ID = 5. The dashed line represents the fitted Gaussian model.

C. Triphasic muscle pattern ($u_{triphasic}$)

The best performance in terms of triphasic muscle pattern $u_{\text{triphasic}}$ across the majority of difficulty indices (ID = 2

TABLE I

Analysis of movement characteristics of each combination of model and task requirement using the proposed metrics[§].

Metric	Task requirements	Models															
		Baseline				Execution noise				Optimality principles				Hybrid			
		Index of difficulty (ID)				Index of difficulty (ID)				Index of difficulty (ID)				Index of difficulty (ID)			
		2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
p_{line}	pos	0.97	0.97	0.97	1.0	0.99	**1.0**	0.97	0.96	0.96	0.95	0.94	0.99	0.98	0.94	0.96	0.98
	pos-vel	0.91	0.95	0.95	0.97	0.99	**1.0**	0.99	0.99	0.96	0.97	0.93	0.94	0.97	0.92	0.99	**1.0**
	pos-vel-acc	**1.0**	0.98	0.98	1.0	0.98	0.99	**1.0**	**1.0**	0.96	0.99	0.96	0.98	0.95	0.93	0.97	0.98
v_{bell}	pos[†]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	pos-vel	0.86	0.73	0.80	0.94	0.81	0.80	0.76	0.90	0.86	0.88	0.82	0.91	0.88	0.83	0.89	0.94
	pos-vel-acc	**0.90**	0.78	0.93	0.95	0.74	0.80	0.79	0.95	0.89	**0.92**	0.81	0.91	**0.90**	**0.92**	**0.95**	**0.97**
$u_{\text{triphasic}}$	pos	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0
	pos-vel	1^*	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	pos-vel-acc	1^*	1	0	1^*	1	1	1	1^*	1	1	1	0	1	1	1	1^*
R_F[‡]	pos	0.985				0.967				0.974				0.967			
	pos-vel	0.956				0.997				**0.998**				0.983			
	pos-vel-acc	0.963				0.985				0.986				0.929			

[§] The highest values between task requirements for each metric, model, and difficulty index are highlighted in green. Among these values, the best performance across difficulty indexes for each metric is highlighted in bold dark green.

[†] The trajectories of this terminal condition are invalid for the v_{bell} metric as they do not reach the lower threshold of 10% of the peak velocity.

[‡] R_F has only value because this metric determines the correlation coefficient across all difficulty indices (ID=2...5).

[*] These combinations exhibit two or three muscle pairs with a triphasic muscle pattern.

to 5) for all models (Baseline, Execution noise, Optimality principles and Hybrid) is linked to pos-vel and pos-vel-acc task requirements (Tab. I). The muscle patterns for ID $= 5$ of hybrid model for each task requirement are shown in Figure 4. The figure illustrates that both pos-vel and pos-vel-acc task requirements give rise to a triphasic muscle pattern in the elbow muscle pair, whereas only position task requirement (pos) results in a biphasic pattern in the three muscle pairs. Furthermore, the figure displays two triphasic muscle patterns for the pos-vel-acc task requirement. Similarly, Figure 5 illustrates that duration of the third muscle phase increases with larger index of difficulty (ID).

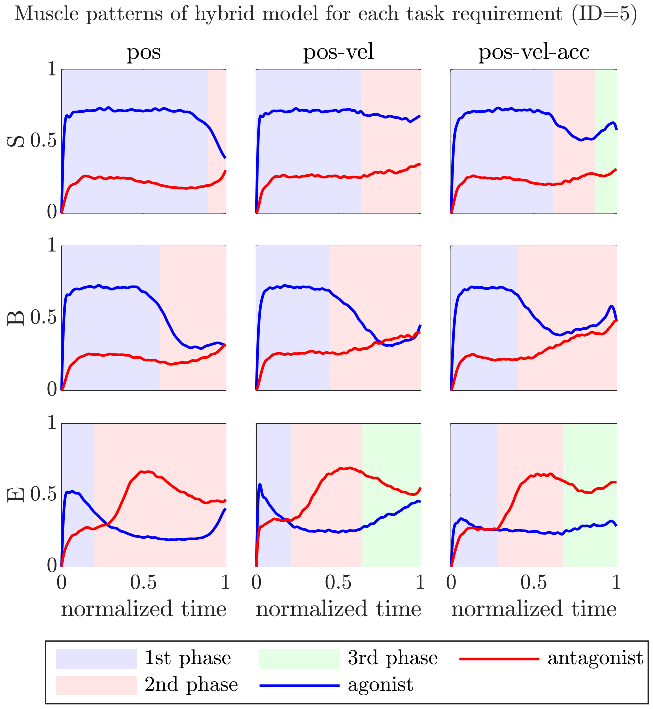

Fig. 4. Muscle activation pattern of the hybrid model, considering the three possible task requirements. The agonist-antagonist muscle pair of the arm are denoted as: Monoarticular shoulder (S), Biarticular elbow-shoulder (B) and Monoarticular elbow muscle (E). Blue and Red lines represent muscle activation of agonist and antagonist muscles, respectively.

D. Fitts's law (R_F)

The models Execution noise, Optimality principles and Hybrid obtain their highest correlation coefficient R_F when incorporating velocity requirement into the main task (pos-vel) (Tab. I). In contrast, the Baseline model attains its highest correlation coefficient R_F when employing only the position task requirement. The optimality principles model demonstrate slightly superior performance in R_F compared to Baseline, Execution noise and Hybrid models. The linear relationship between Movement Time (MT) and Index of Difficulty (ID) is graphically illustrated in Figure 6 for the Optimality Principles model considering the three possible task requirements. Additionally, the figure illustrates the increase in movement time variance as more kinematic requirements are incorporated into the main task. It is crucial to emphasize that all combinations generate a robust linear relationship, with correlation coefficients $R_F > 0.9$. Consequently, although certain combinations exhibit higher correlation coefficients than others, it is implausible to indicate which combination yields the most realistic arm trajectory based solely on the R_F.

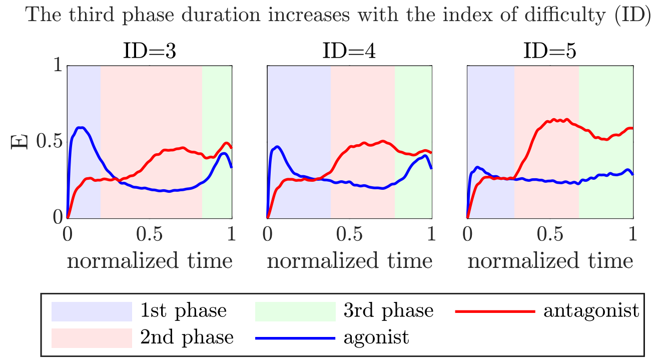

Fig. 5. Muscle pattern of Monoarticular elbow muscle (E) for hybrid model with pos-vel-acc task requirement. The third phase duration increases with larger index of difficulty (ID), i.e., higher endpoint accuracy.

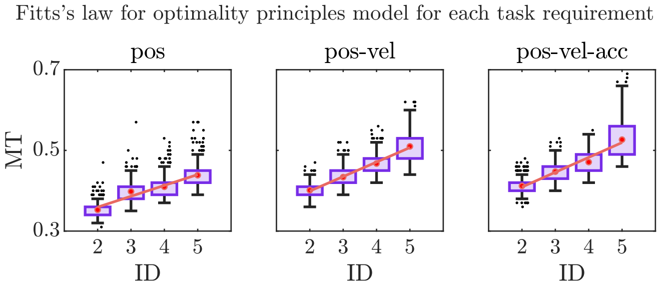

Fig. 6. Graphic representation of the linear relationship between Movement Time (MT) and Index of Difficulty (ID) for the optimality principles model, considering the three possible task requirements. The red dots represent the average movement time.

IV. Discussion

Through the systematic combination of factors (a-c), we identify three key considerations for generating human-like characteristics in point-to-point arm movements. First, including both velocity and acceleration as task requirement ((a), pos-vel-acc), results in good or excellent values across all metrics and in the majority of difficulty indices, regardless of the model. Second, using noise b) during movement execution, in combination with reward terms c) minimizing mechanical work, hand jerk, and control effort results in the most bell-shaped velocity profile across the majority of difficulty indexes (except for the position-only task requirement). Third, a higher endpoint accuracy, i.e., a larger index of difficulty (ID), leads to a longer duration of the third muscle phase and velocity profiles with a better-defined bell shape. According to Wierzbicka et al. [7], the role of this third phase is to regulate the braking forces to guide the hand towards the target position. Therefore, the effect of the index of difficulty (ID) can be understood as the prolongation of the deceleration phase to achieve high endpoint accuracy, which in turn smooths the velocity profiles on the right side, enhancing the bell shape. It is noteworthy that although increasing the index of difficulty (ID) has improved the bell shape, larger values will cause the velocity profile to become more positively asymmetric [21].

In addition, we found that including the velocity requirement into the reaching task (pos-vel) can yield comparable results to considering both velocity and acceleration (pos-vel-acc). The primary distinction lies in slightly lower values of bell-shaped velocity profile v_{bell}. Moreover, we found that the triphasic muscle pattern can emerge when incorporating requirements of either velocity (pos-vel) or velocity and acceleration (pos-vel-acc) into reaching tasks. This contrasts with Ueyama et al. [13], who suggests position, velocity and force (equivalent to acceleration) are necessary. Unlike Ueyama et al. [13], our approach does not require predefining the movement time for arm movement generation. Although it is not clear how predefining the movement time (MT) influences the emergence of the triphasic muscle pattern, setting a value far from that calculated with Fitts's law could result in unrealistic arm movements.

It is noteworthy to highlight that the metrics p_{line}, $u_{\text{triphasic}}$ and R_F do not show large differences across all combinations. This suggests that all investigated models and task requirements (except for position only) lead to somewhat realistic arm movements, at least for the simple planar point-to-point movements investigated here.

Although our control approach generates realistic arm movements with human-like characteristics, our study has some limitations. The arm model used incorporates only two degrees of freedom and six muscles. Consequently, our model does not fully account for the entire joint and muscle redundancy found in a real human arm. Furthermore, the investigated task includes only point-to-point reaching tasks, whereas more openly defined tasks such as point-to-manifold reaching might be interesting for future research, as they offer more freedom in arm movement generation. Previous studies [10], [12] have shown significance differences in the generated arm trajectories using point-to-manifold reaching that have not been observed in point-to-point movements. Moreover, complex movements in a complex arm model may further distinguish between the different combinations such that a solution for predicting realistic human arm movements with RL could aid the development and control of assistive devices.

References

[1] A. Cieza, K. Causey, K. Kamenov, S. W. Hanson, S. Chatterji, and T. Vos, "Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10267, pp. 2006–2017, 2020.

[2] P. Maciejasz, J. Eschweiler, K. Gerlach-Hahn, A. Jansen-Troy, and S. Leonhardt, "A survey on robotic devices for upper limb rehabilitation," *Journal of neuroengineering and rehabilitation*, vol. 11, no. 1, pp. 1–29, 2014.

[3] P. Schumacher, T. Geijtenbeek, V. Caggiano, V. Kumar, S. Schmitt, G. Martius, and D. F. Haeufle, "Natural and robust walking using reinforcement learning without demonstrations in high-dimensional musculoskeletal models," *arXiv preprint arXiv:2309.02976*, 2023.

[4] F. Fischer, M. Bachinski, M. Klar, A. Fleig, and J. Müller, "Reinforcement learning control of a biomechanical model of the upper extremity," *Scientific Reports*, vol. 11, no. 1, p. 14445, 2021.

[5] P. Morasso, "Spatial control of arm movements," *Experimental brain research*, vol. 42, no. 2, pp. 223–227, 1981.

[6] W. Abend, E. Bizzi, and P. Morasso, "Human arm trajectory formation." *Brain: a journal of neurology*, vol. 105, no. Pt 2, pp. 331–348, 1982.

[7] M. M. Wierzbicka, A. W. Wiegner, and B. T. Shahani, “Role of agonist and antagonist muscles in fast arm movements in man,” *Experimental Brain Research*, vol. 63, pp. 331–340, 1986.

[8] D. A. Kistemaker, A. K. J. Van Soest, and M. F. Bobbert, “Is equilibrium point control feasible for fast goal-directed single-joint movements?” *Journal of Neurophysiology*, vol. 95, no. 5, pp. 2898–2912, 2006.

[9] I. S. MacKenzie, “A note on the information-theoretic basis for fitts' law,” *Journal of motor behavior*, vol. 21, no. 3, pp. 323–330, 1989.

[10] B. Berret, E. Chiovetto, F. Nori, and T. Pozzo, “Evidence for composite cost functions in arm movement planning: an inverse optimal control approach,” *PLoS computational biology*, vol. 7, no. 10, p. e1002183, 2011.

[11] T. Flash and N. Hogan, “The coordination of arm movements: an experimentally confirmed mathematical model,” *Journal of neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985.

[12] I. Wochner, D. Driess, H. Zimmermann, D. F. Haeufle, M. Toussaint, and S. Schmitt, “Optimality principles in human point-to-manifold reaching accounting for muscle dynamics,” *Frontiers in computational neuroscience*, vol. 14, p. 38, 2020.

[13] Y. Ueyama, “Costs of position, velocity, and force requirements in optimal control induce triphasic muscle activation during reaching movement,” *Scientific Reports*, vol. 11, no. 1, p. 16815, 2021.

[14] A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller, “Maximum a posteriori policy optimisation,” in *International Conference on Learning Representations*, 2018.

[15] P. Schumacher, D. Haeufle, D. Büchler, S. Schmitt, and G. Martius, “DEP-RL: Embodied exploration for reinforcement learning in overactuated and musculoskeletal systems,” in *The Eleventh International Conference on Learning Representations*, 2023.

[16] F. Pardo, “Tonic: A deep reinforcement learning library for fast prototyping and benchmarking,” *arXiv preprint arXiv:2011.07537*, 2020.

[17] R. J. Van Beers, P. Haggard, and D. M. Wolpert, “The role of execution noise in movement variability,” *Journal of neurophysiology*, vol. 91, no. 2, pp. 1050–1063, 2004.

[18] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[19] U. Jönsson, C. Trygger, and P. Ögren, “Optimal control-lecture notes,” *Cited on*, p. 15, 2010.

[20] C. M. Harris and D. M. Wolpert, “Signal-dependent noise determines motor planning,” *Nature*, vol. 394, no. 6695, pp. 780–784, 1998.

[21] E. Todorov, “Optimality principles in sensorimotor control,” *Nature neuroscience*, vol. 7, no. 9, pp. 907–915, 2004.