Human-Aware Design for Transferring Knowledge During Human-AI Co-Learning

Dimtrios Koutrintzes[1], Christos Spatharis[1] and Maria Dagioglou[1]

[1]*Institute of Informatics and Telecommunications, National Centre for Scientific Research 'Demokritos'*

Abstract

State-of-the-art AI methods allow us to develop agents that collaborate and co-learn with humans. The possibility to transfer knowledge from an expert to a novice human-AI team has the potential to streamline training, increase productivity and foster a more effective collaborative environment where individuals build on each other's strength. In this context, we present an experimentation pipeline that can be followed during human-aware AI design and development in the case of transfer learning from expert to novice human-AI teams. Moreover, we tackle two intricate research questions of 'when to stop training' and 'what expert knowledge' to transfer. Our results of a study with two expert human participants demonstrate the complexities of process and offer relevant guidlines for future research.

Keywords

Human-AI collaboration, Human-AI co-learning, deep Reinforcement Learning, Transfer learning, Human-aware design, Expert's behaviour,

1. Introduction

Industry 5.0 brings forward a social integration of technology into the factory floor [1]. Humans and society at large come at the centre of artificial intelligence (AI) systems, across their entire life-cycle (from design to deployment and maintenance) through values-driven design, the satisfaction of principles of ethical and trustworthy AI and importantly through cultivating a congruous mentality among the stakeholders (developers, integrators, regulators, etc.). Furthermore, human-centric digitisation challenges us to re-imagine and redesign industrial tasks in a way that human and artificial agency are interwoven into a sustainable and resilient fabric.

'Human-AI collaboration' (HAIC) is an increasingly popular term, describing many different things, and possibly shifting our attention from several ethical issues related to the integration of AI in our society [2]. In the present work, HAIC, similarly to human-robot collaboration (HRC), is used to describe systems where humans and AI (embodied or not) share goals and *perform interdependent actions* and is a different paradigm compared to co-existence, interaction and cooperation [3]. AI collaborators, from games [4, 5] to robots in industrial set ups [6], need to incorporate qualities and capabilities that support fluent and seamless collaboration. Similar to human joint action [7], AI agents need to support processes for common perceptual and cognitive grounding, transparent agency attribution and co-learning [8, 9, 10, 11].

HAII5.0: Embracing Human-Aware AI in Industry 5.0, Workshop at ECAI 2024, Oct. 19, 2024, Santiago de Compostela, ES

dkoutrintzes@iit.demokritos.gr (D. Koutrintzes); cspatharis@iit.demokritos.gr (C. Spatharis); mdagiogl@iit.demokritos.gr (M. Dagioglou)

0009-0003-7401-6347 (D. Koutrintzes); 0009-0001-2791-2291 (C. Spatharis); 0000-0002-3357-2844 (M. Dagioglou)

The successful development of agents that collaborate with humans depends both on the performance of state-of-the-art methods and the study of human behaviour. Deep reinforcement learning (dRL) methods have allowed to develop agents that collaborate and co-learn in real-time and real-world [12, 13]. During the collaboration, it becomes possible to study how humans perceive their interaction with an agent and how they behave in this context.

The present work is related to the study of agents' capabilities for co-learning and transfer learning (TL). Like human de novo learning [14], co-learning demands long training periods and involves a considerable physical effort and cognitive load. Transferring knowledge from *expert* human-AI teams (HAIT) to *novice* HAITs can alleviate these complexities and support retaining expert knowledge [15]. In HAIC, it is possible that the source and target of TL are human individuals, while the environment remains constant. This means that TL is not between different tasks or settings, but between people working in the same context. The expertise, skills or insights of one person can be used to accelerate the learning and performance of another. This 'person-to-person TL' has the potential to streamline training, increase productivity and foster a more effective collaborative environment where individuals build on each other's strengths [15].

The process of transferring knowledge comprises several challenges that can be perhaps formulated as a trade-off between transferring the knowledge to *perform* and transferring the knowledge to *learn*. Time and cost efficiency reasons might require opting for performance. On the other hand, learning to learn leaves more space to individualisation, avoids 'experts-biases' and can lead to more sustainable and resilient behaviours in the long run, especially in tasks that involve motor learning [16]. To address such research questions we need adequate HAIC studies that explore different TL techniques [17] and evaluate both the performance of HAITs in the collaborative task, as well as individualisation margins and subjective human attitudes.

Beyond different TL methods, within the context of each method there are many design and development considerations during the stage of training an AI agent with the expert human. There is a number of choices, seemingly technical, from model initialisation to deciding when to stop the training that can impact knowledge transfer and merit investigation on their own.

In this context, in the present work we explore the training process of AI agents with expert humans, as well as with trained expert agents, and demonstrate the complexity of the design choices at hand. Our main research question is about "How to evaluate behaviour and when to stop training an expert AI agent and transfer this knowledge to novice HAITs (for further behavioural studies)?". We provide design considerations that allow human-aware TL during human-AI co-learning. We then present the results of a study with two expert human participants that demonstrate the complexities of deciding what knowledge to transfer and when to transfer it.

The rest of the paper is organised as follows. Section 2 presents the work related to the present paper. Section 3 describes the methods of the present work, including design considerations for human-aware transfer learning in human-AI co-learning, the used co-learning task, the details of the AI agents, as well as the experimental design and conditions. Section 4 reports the related results. Finally, Section 5 discusses our findings and Section 6 concludes this report with future challenges and research directions.

2. Background

Recent advancements in deep reinforcement learning (deep RL) have enabled the deployment of complex systems that operate in real-time in dynamic environments. The success of deep RL arises from its ability to learn complex motions and behaviours that are challenging to generate using traditional hard-coded solutions. For example, dRL has been successfully implemented for various robotic capabilities such as: navigation [18], robot arm control [19], grasping [20], drones maneuvering [21] and human-robot co-learning [12]. One popular paradigm in co-learning tasks is Soft Actor-Critic (SAC) [22] due to its robustness in balancing exploration and exploitation, which is crucial in interactive environments [12, 23, 13].

Recently, there has been a growing emphasis on developing agents that engage with humans. Several studies focus on games [24, 25, 26], however, due to their competitive nature, these agents typically rely on choosing the best actions against a nearly optimal opponent. Conversely, in collaborative or social settings, modelling and leveraging human behaviour to work along agents is a highly challenging task [27]. In the context of human-AI co-learning, both entities can learn from each other and grow together over time[28]. It must be noted that most HAIC studies operate within well-defined discrete environments, such as *overcooked* [27]. In contrast, our work addresses a continuous environment that necessitates collaboration between humans and agents to generate multi-modal trajectories and collectively achieve a goal.

A challenge of deploying dRL models in diverse or previously unexplored environments [29] is to do so without requiring training from scratch. Especially in HAIC tasks, this process is time-consuming and demanding, and the extended training periods can negatively affect the performance of the team. These limitations highlight the importance of a different paradigm that allows agents to reuse knowledge from one task to a related, yet distinct, task. In RL tasks, this paradigm is referred to as transfer learning (TL).

Multiple approaches have been proposed to facilitate knowledge transfer in RL settings [30, 31]. TL aims to learn an optimal policy for a target task by leveraging external information from a set of source tasks, as well as, internal information from the target task. One of the most prominent TL approaches, *reward shaping (RS)* [32, 33], uses external knowledge to modify the reward distribution in the target task, by incorporating a reward-shaping function. By providing additional rewards along with interior environmental rewards, RS directs the agent toward more optimal trajectories. *Learning from demonstrations (LfD)* allows RL agents to learn to perform tasks by observing expert demonstrationsThis approach can be further decomposed to offline and online LfD, where the former uses demonstrations for pre-training the models [34, 35], while the latter directly employs expert demonstrations to guide the agent's behaviour for more efficient exploration [36]. Finally, in *policy transfer*, a pre-trained policy on a source task is directly applied to the target task. Policy transfer is further divided into *TL learning via policy distillation* [37] and *TL via policy reuse* [38]. Policy distillation involves learning a model by minimizing the divergence from multiple expert policies, while policy reuse leverages previously learned policies by allowing the agent to draw from past experiences with some probability.

Transferring knowledge in the context of HAIC comprises several complexities. Given the nature of human (motor) learning, a core question is when to stop expert training and to transfer knowledge. Just converging to a desired performance might not be the pursued goal. Injecting variability in the transferred knowledge might be necessary to promote learning and leverage future behaviour [16].

The human involvement in HAIC tasks demands a considerable time and effort from *human experts* due to their active involvement throughout the entire agent training process. This significantly constraints the fine tuning of dRL model hyperparameters. Various studies [39, 40, 41] offer valuable insights on the selection of hyperparameters based on the methodologies and environmental contexts. However, the assumption of hyperparameter tuning does not directly apply to HAIC tasks, as the inclusion of humans in the training loop renders exhaustive hyperparameter search infeasible, given the time and energy constraints involved.

Finally, every hypothesis on the efficiency of a TL method needs to be evaluated through its impact to the entire team, taking into account multiple teams. Evaluating the performance of HAIC teams requires a holistic examination of the human-AI team performance and co-learning dynamics during the collaboration, as well as analyzing individual contributions [42, 43]. Moreover, both objective and subjective metrics need to be incorporated in the evaluation process to comprehensively assess the effectiveness of the team, as well as individual human behaviours and experiences [44].

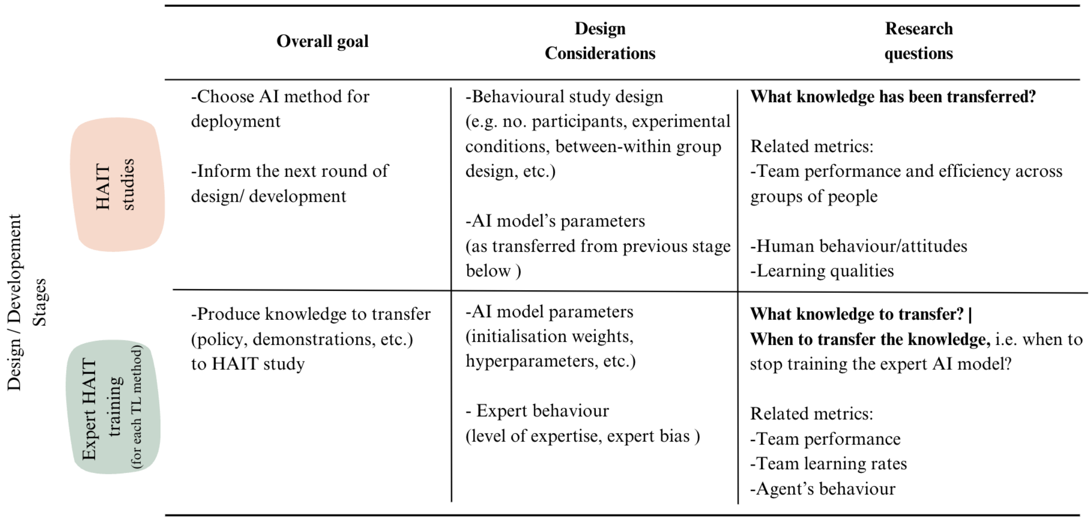
3. Methods

3.1. Design considerations for human-aware transfer learning in human-AI co-learning

The ultimate goal of this work is to build AI agents that possess human-AI co-learning capabilities. Transferring knowledge from expert HAITs can facilitate reasonable training periods for a novice HAIT. Different TL methods are expected to result in different HAIT behaviour and affect human behaviour and perceived interaction qualities. Which TL method allows faster or more stable learning in the long run? Which TL method promotes individualisation and alleviates superstitious learning as a result of expert-behaviour bias? These are examples of research questions that can be pursued through rigorous testing during human studies that attempt to capture 'what knowledge has been transferred'. Within the context of TL, we need to consider two design/development stages. These are presented in the table of Figure 1 and capture our experience during HAIT studies ([13, 45] and other unpublished data). The overall goal of HAIT studies is to either inform the next round of design and development or to actually choose a deployable system (first row of the table in Figure 1). Experimental design and AI model parameters need to be considered.

In the case of TL methods, there is another experimental stage of design/development that precedes that of HAIT studies. This stage is related to training expert HAIT teams (or possibly expert AI-only teams) and aims at producing the knowledge to be transferred (policy, demonstrations, etc.). Any design choices here will determine the AI agent's model parameters in the HAIT studies. In a sense, this is a set of design considerations, besides behavioural experimental design, that needs to be controlled. Rigorous design and reporting of this stage is necessary for guaranteeing transparency of the methods and reproducibility of the results, as well as, for facilitating comparison among studies and methods. We identify two main complexities, as a result of having a human in the loop.

Effort of choosing AI model's hyperparameters. In the absence of humans, iterative



Design / Developement Stages	Overall goal	Design Considerations	Research questions
HAIT studies	-Choose AI method for deployment -Inform the next round of design/ development	-Behavioural study design (e.g. no. participants, experimental conditions, between-within group design, etc.) -AI model's parameters (as transferred from previous stage below)	**What knowledge has been transferred?** Related metrics: -Team performance and efficiency across groups of people -Human behaviour/attitudes -Learning qualities
Expert HAIT training (for each TL method)	-Produce knowledge to transfer (policy, demonstrations, etc.) to HAIT study	-AI model parameters (initialisation weights, hyperparameters, etc.) - Expert behaviour (level of expertise, expert bias)	**What knowledge to transfer? |** **When to transfer the knowledge,** i.e. when to stop training the expert AI model? Related metrics: -Team performance -Team learning rates -Agent's behaviour

Figure 1: Design considerations for human-aware transfer learning in human-AI co-learning.

testing of different sets of hyperparameters is exploited until a desirable performance is achieved. In the event that human experts need to train the models, exhaustively investigating appropriate parameters requires enormous effort and time. Alternative approaches such as using two independent agents (instead of an agent and a human) for team training could be pursued. However, based on our experience this might not work due to lack of representing important task aspects (such as the collaborative nature of a task). On the other hand, the use of collaborative agents introduces issues of decoupling the information later on.

When to stop the training? While in a 'AI-alone' system the goal is to converge to the best possible performance, in the case of transferring knowledge for co-learning this might not be the desired outcome. Instead, a certain degree of variance in the transferred knowledge might be pursued as a means of facilitating individualisation. As mentioned earlier, individualisation prevents 'experts-biases' and can lead to more sustainable and resilient behaviours in the long run. Such a design choice will also affect the set of chosen parameters.

In the rest of the paper we focus on the 'expert HAIT training' and we demonstrate through an experimental paradigm the process and results towards defining the knowledge to be transferred before proceeding to a HAIT study.

3.2. Human-AI co-learning task

A co-learning task, in a virtual environment, has been used to study HAIT behaviour, and specifically expert HAIT training [23]. A human collaborates with an RL agent to move a ball from a starting to a target position (Figure 2), along a virtual tray. The tray rotates around two axes; the human player controls the rotation around the y-axis, while the agent controls the rotation around x-axis. The controlled variables are the angles θ, φ of the tray's rotation. A ‘game’ lasts for maximum 40 seconds and the team wins if the ball reaches the target before this. A pair of human and agent actions is applied for 200ms resulting in 200 time-steps per game.

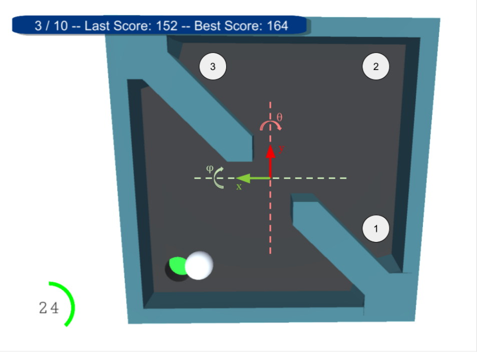


Figure 2: HAIC task virtual environment. A white ball (d=1-unit) travels from one of three starting positions (enumerated white circles) to a target position (green hole), along a 10x10 unit tray. The ball is constrained within the tray by a 1-unit high wall. Two obstacles placed across the main diagonal of the tray force the ball to move through through a 1.4-unit ‘gate’ along its trip from the start to the goal positions. A count down timer is presented to users during a game (left hand side bottom), while study and time statistics are shown at the end of each game.

Both team members can take three discrete actions: a) rotate the tray clockwise, b) rotate the tray counter-clockwise, or c) leave tray’s angle unchanged. The human collaborator applies these actions through the keyboard by pressing ‘Right Arrow’ (>), ‘Left Arrow’ (<) or nothing, accordingly. The tray rotates 30 degrees towards both sides, and each action causes an angle change of around 5 degrees.

3.3. Deep RL agent

The rotation of the tray around x-axis is controlled by an AI agent. A discrete version of Soft-Actor Critic (SAC) has been used [46] to be more consistent with the discrete inputs provided by the human user via a keyboard. A continuous 8-dimensional state space has been designed to represent the environment’s configuration at each time-step, comprising: the ball’s position (x, y) and speed $(\dot{x}, \dot{y})$ and the tray’s angles (ϕ, θ) and rotational velocities $(\dot{\phi}, \dot{\theta})$.

The above quantities have been normalized to the $[-1, 1]$ range to ensure training stability for the neural networks. The dRL agent’s action space is 1-dimensional and discrete, $\alpha = -1, 0, 1$, corresponding to counter-clockwise, no or clockwise change of the tray’s angles accordingly. The agent receives a $r = -1$ reward for each elapsed time-step, and an $r = 10$ reward when the team reaches the target. Figure 3 depicts the co-learning loop between the human and the RL agent.

3.4. Human-AI co-learning process

Two experts were involved in the training process. They are regarded as experts due to their profound understanding of the game’s environment, and agent’s behaviour. Their expertise has

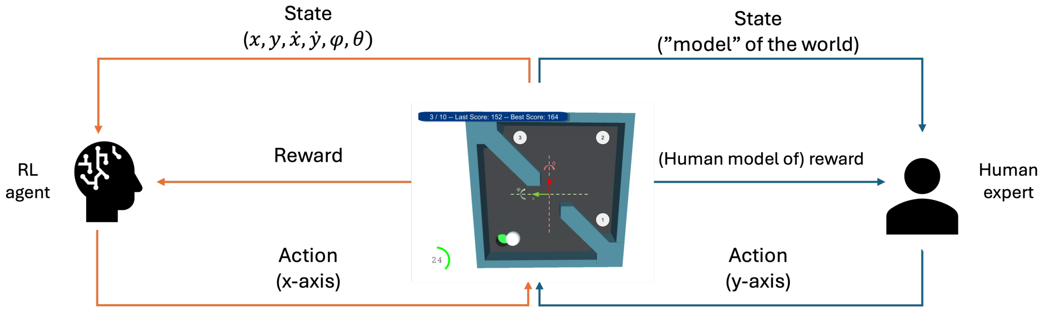


Figure 3: Co-learning process between an RL agent and a human expert.

been developed through extensive gameplay, with each expert having spent a minimum of 100 hours playing the game.

The human-AI co-learning process is presented in Figure 4. Each team had to complete six experimental blocks, where each block comprised:

-A **testing batch** of 5 games, where the agent followed a deterministic approach by applying the argmax action of its policy. No interaction data were collected for the replay buffer during this phase.

-A **training batch** of 5 games, each of which was followed by a round of 250 off-line gradient updates (OGU). Here, the policy followed a stochastic approach by sampling from the Actor's categorical distribution. Interactions were stored in a buffer and used for OGU after each game. Training began after the third game of the first block to ensure that the buffer had enough data. So, in total there were 7000 OGU across the entire learning procedure.

At the end of the sixth block, one more testing batch was included. For all experiments, and for both experts, the same initialization weights have been used. This was motivated by the variability in the subsequent performance that this initial agent induces. The initialization weights were selected randomly without any evaluation of their impact to the subsequent team performance. None of the expert had any previous experience with this initialized agent.

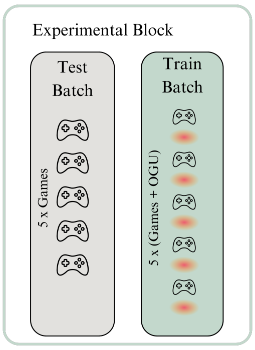


Figure 4: An experimental block consists of 5 test games (test batch) and 5 training games (train batch). Each training game is followed by an off-line gradient update (OGU).

Each human expert repeated the co-learning procedure three times under two conditions. In

Experiment 1 the humans collaborated with a SAC agent. This has been the *baseline condition* that was used during hyperparameter tuning. Observations during this experimental condition are exploited to define the final set of hyperparameters as well as the expert HAITs behaviour to be transferred. Note that the number of the experimental blocks (6) was chosen after observing the convergence of performance during experimentation in the condition of Experiment 1. While behaviours of individual experts could be certainly transferred, we wanted to explore the effect of combing experts' knowledge as a means of considering various, potentially different, expert behaviours. In **Experiment 2** the initialised agents were pre-trained using a combination of two replay buffers sourced from the two experts (during Experiment 1). Each expert selected their replay buffer based on their assessment of the best run. During the pre-training phase, the agent underwent 2500 gradient updates. The final replay buffer contained approximately 3500 interactions and new interactions during the collaboration replaced the old ones. In addition to human-AI teams, in **Experiment 3**, we subjected to the co-learning procedure three pairs of independent agents. Each pair consisted of a 'novice' SAC agent and a SAC agent as pre-trained by the experts during Experiment 1 (direct policy transfer).

The performance of the teams across the games is evaluated by the achieved score in each game. This is computed by discounting one point for each time-step played per game. For example, a successful game of 10 seconds (50 time-steps) would result in a score of 150, while a unsuccessful game would result in a score of 0. Moreover, the performance is qualitatively evaluated through occupancy grids of the ball throughout the games.

4. Results

4.1. Hyperparameter tuning

The hyperparameter values reported in Table 1 were defined after a series of experiments to optimize the performance of the agent for the HAIC task. This time-consuming process involved continuous human interaction in the training loop, requiring iterative testing and validation. The involvement of human-in-the-loop not only slowed down the cycle, but also introduced variability, necessitating numerous trials to converge to optimal settings.

Table 1
Final set of hyperparameters.

Hyperparameter	Value
Number of layers	2 fully connected layers
Hidden layer units	[32, 32]
Replay buffer size	3500
Off-line gradient updates	250
Batch size	256
Discount rate	0.99
Learning rate	0.0003
Optimizer	Adam
Weight initializer	Xavier
Target entropy	0.5 * (-log(1/|A|))

Following, we discuss the various hyperparameters adjusted during training and analyze their impact on the overall performance of the method.

Target entropy The target entropy $\overline{H}$ is crucial in SAC as it balances exploration and exploitation. The equation to calculate the optimal target entropy contains a multiplier and the formula that maximizes the entropy to give all actions the same probability. A high multiplier maximizes the exploration but makes the agent having a more random behaviour, while a low multiplier allows more intense exploitation with minimum exploration. Based on our experience, using multipliers based on other set-ups [46] does not work and experimentation is needed with different multipliers [47] considering the context and the goal of each task. In our case, we pursued a balance between exploration and exploitation considering a specific training time and the desired variability in the experts' behaviour.

Buffer sizes. The replay buffer (RB) is used to store past experiences to update the policy in SAC. We tested various buffer sizes and our findings align with previous research [48], showing that small RB sizes discard useful experiences over time, while very large sizes can also negatively affect performance by including outdated and irrelevant experiences. Considering that in each block the maximum number of experiences we can collect is 1000 (200 timesteps x 5 games per block), we opted to use a replay buffer size of 3500 experiences. This size allows us to discard initial sub-optimal experiences (from the first two blocks) and keep the latest ones, ensuring that as the policy progresses, the co-learning process prioritizes the most recent experiences gathered.

Update frequencies. A significant effort was dedicated on identifying the optimal frequency for performing offline gradient updates (OGU) on the neural networks using experiences stored in the replay buffer. According to our findings, performing multiple OGUs after many rounds of interaction (e.g. after each block) between the human and the RL agent might result in stagnant policies that contribute to a poorly filled replay buffer, which will not be as effective for the update process. On the other hand, updating very frequently (e.g. during the game), while the RL agent actively gathers experiences, could end up with the opposite result: the policy changes while playing and confuses the co-learning process. In our case, executing multiple OGUs at the end of each game achieved a good balance between policy exploitation and updating, hence we believe that the update frequency depends greatly on the pipeline of the overall methodology.

AI-only training. In an attempt to further optimize the hyperparameters, we considered experimenting without human involvement in the training loop. To that end, we implemented an agent-vs-agent training scheme, where one agent controlled one axis of the tray while another controlled the other axis. The goal was to identify an optimal set of hyperparameters through multiple experiments that could be conducted seamlessly in the absence of the human expert. However, the independent agents failed to learn the task, highlighting the complexity of hyperparameter tuning in multi-agent systems and the irreplaceable value of human expertise in collaborative tasks.

Overall, hyperparameter tuning in such tasks is particularly challenging as human involvement is required. Additionally, the human must decide when to stop testing a set of hyperparameters and move on to a new one, making the process even more complex. Despite the difficulties, balancing exploration and exploitation and leveraging human expertise are essential for effective training.

4.2. Experiment 1: Human - SAC agent co-learning

Two human experts collaborated with the AI system (SAC agent) from the scratch. No prior experience was used in the replay buffer. Figure 5 (left) shows that both experts exhibit comparable behaviour, converging to a high score towards the last testing blocks as a manifestation of expert behaviour. On the other hand, blocks 1 to 4 show significant variability in the performance of both experts, suggesting that in the initial stages of training, the human-AI team still explores various strategies to jointly reach the target.

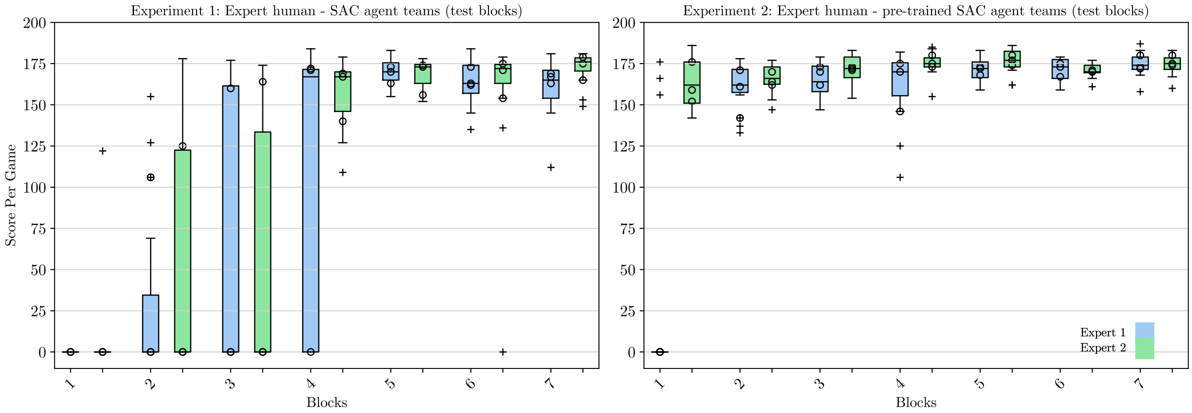


Figure 5: Scores of HAIT during the test games for Experiment 1 (left) and Experiment 2 (right). For each expert, the games across all three runs are collapsed in each boxplot. The median for each run is shown by the open black circles.

Figure 6 shows the ball's occupancy frequencies across the test blocks. In the first block, the team performs poorly, with the ball remaining in the upper part of the tray, where cells are highly occupied. In contrast, the second block shows more exploratory behavior, as the team searches through the tray in order to find ways to reach the goal. As testing blocks proceed, the coverage becomes sparser, indicating that HAIT has converged towards a more direct approach to reaching the target, demonstrated by the 'X'-shaped occupancy grid, with increased occupancy in cells near the target, such as the lower left cell.

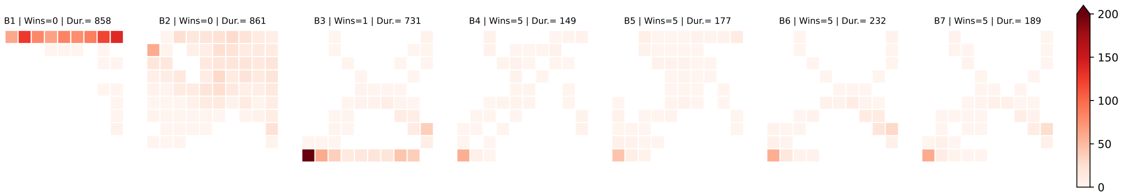


Figure 6: Ball occupancy frequency across the seven blocks during the test batches of one HAIT in Experiment 1.

4.3. Experiment 2: Human - Pre-trained SAC agent co-learning

During the second experiment we wanted to study the effect of TL within experts, as a possible means of introducing some variability due to the different experts. A replay buffer that combined knowledge from the two replay buffers of two experts, during Experiment 1, was used. These data were used to pre-train the SAC agent before interacting with the human experts [36]. Specifically, we conducted 2500 OGUs prior to starting the games, and then followed the same experimental setup as in Experiment 1.

The results, depicted in Figure 5 (right), demonstrate the effectiveness of TL in the collaborative task. Despite some poor performance (of one of the two HAITs) in the very first test block the overall performance of the teams across the rest of the block is consistently high. This shows the robustness and the capability of offline TL scheme to produce high quality solutions. The observed variability in the initial blocks can be attributed to both variance in human performance but also to some limited continuation of learning as can be seen in the heatmaps of Figure 7 (Blocks 1 and 2) .

Furthermore, in Figure 7, it can be noticed that the evolution of the occupancy grids differs from Experiment 1. In particular, from the very first test block, the HAIT tends to explore the lower half of the grid, while frequently reaching the lower left cell near the goal. In the following test blocks, the team appears to quickly achieve optimal behaviour (i.e., ‘X’-shaped occupancy grid), successfully reaching the goal from any starting position. This knowledge reuse approach essentially continues the learning process from where it concluded at the end of Experiment 1 and the variability introduced in the pre-training procedure (joined expert buffers) due to possibly different expert behaviours is not sufficient to trigger a significant drop in the initial HAIT performance.

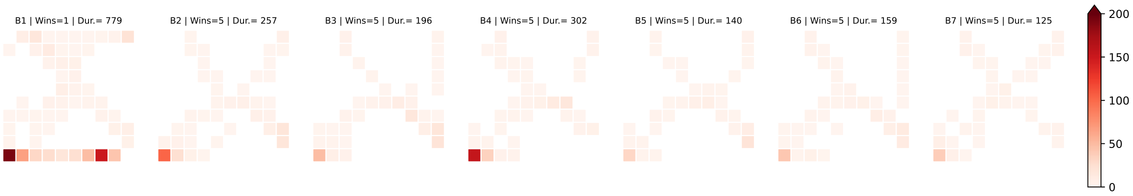


Figure 7: Ball occupancy frequency across the seven blocks during test batches of one HAIT in Experiment 2.

4.4. Experiment 3: Pre-trained - novice SAC agents co-learning

In a final set of experiments, we employed TL with a focus on direct policy transfer to observe the performance of two agents collaborating to achieve a common goal. Specifically, one agent was initialized with a pre-trained policy from Experiment 1, while the other agent started with no prior experience. Only the second agent participated in the training process, while the parameters of the pre-trained agent remained fixed throughout the entire procedure. The pre-trained agent was frozen to retain the behaviour learned by the experts, and better simulating in this way the condition of the previous experiments where we consider that the behaviour of experts humans has reached a certain plateau.

Figure 8 shows the scores obtained during the test blocks. Despite the first agent being equipped with a (sub-)optimal expert policy, the overall team performance was poor compared to the previous experiments. Specifically, the agents failed to achieve the high scores seen previously and exhibited high variability, indicating that the team had not developed the necessary collaborative skills. This further highlights the crucial role of having a human expert in the loop.

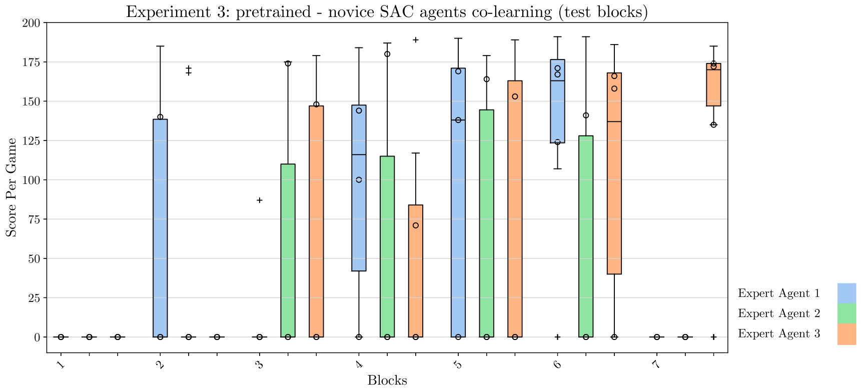


Figure 8: Scores of agent teams during the test games for Experiment 3.

As anticipated, the grids shown in Figure 9 fail to demonstrate meaningful behaviours, as the team is rarely able to successfully reach the target. Instead, there is a noticeable tendency for the agents to remain stuck along the edges of the grid for extended time-steps.

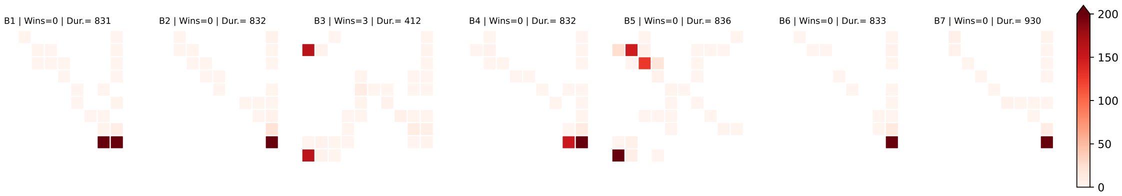


Figure 9: Ball occupancy frequency across seven blocks during the test batches of one run in Experiment 3.

Additionally, we considered another evaluation scheme between a pre-trained and a novice SAC agent. Specifically, we pre-trained one agent offline using batches from the replay buffer that contained experiences from both experts, similar to Experiment 2. Following the same procedure as in Experiment 3, the pre-trained agent had its weights frozen while the novice agent underwent training. The results exhibited the same behaviour as the previous one, demonstrating poor performance, further supporting the need for a human expert in the training procedure. Furthermore all AI-only schemes were tested with the pre-trained agent participating in the training process but produced similar or worse results.

5. Discussion

Recent advances in AI, such as in dRL, allow us to develop systems where humans and AI agents/robots learn together and collaborate to achieve common goals in scenarios where their actions are interdependent. The design, development and validation of 'human-AI collaboration' (HAIC) systems comprises not only the development of the AI methods but also a vigorous study of 'what works for human collaborators' and for the human-AI team (HAIT) at large. Such an approach, has been inherent to fields such as human-robot interaction but is now widely appreciated in the context of AI ethical assessment processes and the human-centric design elements required by Industry 5.0.

Along with the capability for collaboration comes also the necessity of developing methods that allow HAITs to co-learn. Although the technology to support this exists, learning is a long process by nature. The possibility to transfer knowledge from an expert HAIT to a novice HAIT could shorten training periods, increase productivity and prevent loss of expert knowledge.

In the present work, we have first listed several considerations for designing, developing and deploying human-AI co-learning systems. These considerations come out of our experience and follow practices of human-aware AI design. One of the most important aspects of having humans-in-the-loop is that any 'final' solution needs to be validated with many users in order to evaluate not only the suitability of TL methods chosen but also the entire collaborative process as experienced by humans. This is already a complicated procedure that needs careful and controlled experimental designs due to the very nature of humans that exhibits great variability. Moreover, the execution of a HAIT studies for TL, presupposes that expert HAIT knowledge has been captured (one method or another) and what is then evaluated is the effect of the transferred knowledge in the co-learning process.

Based on our experience, an important step that is necessary before any HAIT study for TL from expert HAITs is related to the very procedure of 'knowledge collection' from expert HAITs. A major complexity is related to choosing appropriate hyperparameters for the AI models as the human-in-the-loop nature of HAIC makes hyperparameter fine-tuning a costly procedure. As shown in Experiment 3 (Section 4.4), it could be the case that exploring suitable hyperparameters through AI-AI co-learning might not be possible as was not in our case. The chosen hyperparameters in Section 4.1 have been the result of tens of hours of game training that involved the expert human players. As mentioned earlier each expert has spent over 100 hours of training. This means that the exploration of hyperparameters is constrained both by the effort needed by each individual expert, but also by the fact that a few experts might be available.

In this context, method designers and developers need to decide what constitutes a 'satisfactory behaviour' for a given task and context, and terminate the exploration of hyperparameters based on tailored criteria rather than an optimal performance. This has been the question that we pursued in the presented work: "when to stop training the expert agents and transfer the knowledge to novice teams of humans and experts?". The choice of hyperparameters in Section 4.1 and the results of Experiment 1 (Section 4.2) actually mirror our choices for stopping the training procedure. Two important criteria for doing so are related to the characteristics of the learning curve during the HAIT co-learning and the duration of the procedure that will affect the time required for each participant in HAIT studies later on. Specifically, having in mind that we want to study the effect of TL in novice HAITs using a learning from demonstrations approach, the hyperparameters were chosen so as to:

- have a learning curve that is neither steep nor shallow (regulated by the target entropy). Such a curve allows the final buffers to include demonstrations that mirror the entire learning process, including both bad and good games.
- exclude from the buffers the very initial games that had sub-optimal experiences.
- not intervene in an obtrusive and destructive way in the learning process by inappropriate frequency of the off-line gradient updates.
- not exceed 6 experimental blocks in the future HAIT studies with novice HAITS.

Generally, in terms of "what knowledge to transfer" that is related to the TL method used is related to the follwoing possibilities:

- Transfer knowledge from optimal performance towards the end of learning. This approach could aid the novice players receive refined strategies, potentially leading to quicker adaptation to expert-level behaviors and higher performance.
- Transfer knowledge from an earlier stage where greater variability exists and which could possibly allow more individualisation to the behaviour of novice users. This approach could be more flexible and adaptable to different users' needs, preferences, and learning styles.
- Combine the knowledge of two or more experts, which could also provide some source of variability in the behaviour. By integrating diverse expert experiences, novice players could benefit from a richer set of policies potentially leading to faster convergence as the RL agent has explored the state space more deeply. Note that such variability was shown to leave experts' behaviour unaffected (Section 4.3).

As a final note, we believe that the co-learning paradigm presented satisfies the needs of an experimental set-up. Results produced in such environments can definitely guide design and development in other contexts and tasks, as well as to inform human-aware AI design. However, the design of each system must be treated uniquely based on the specific characteristics of each environment and the participating actors.

6. Conclusions

In the present work, we have demonstrated the complex dynamics involved in developing agents capable of collaborating and co-learning with human experts. Specifically, we have presented an experimentation pipeline that can be followed during human-aware AI design in the case of transfer learning from expert to novice HAITs. Moreover, we tackled two intricate research questions of 'when to stop training' and 'what expert knowledge to transfer'. Through reporting the results of the process we followed we aim at contributing to future research designs that are according to the needs of Industry 5.0 and trustworthy AI.

The next step in our research involves examining how the choices outlined above affect the transfer of knowledge to novice HAITs. Future studies will focus on assessing human behavior and subjective perceptions of collaboration in human-AI interactions, in addition to objective team performance. By evaluating the transfer learning capabilities of our method with novice human HAITs, we aim to validate our findings and further refine our approach.

Acknowledgments

This research was (co-)funded by the European Union under GA no. 101135782 (MANOLO project). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or CNECT. Neither the European Union nor CNECT can be held responsible for them.

References

[1] E. Commission, D.-G. for Research, Innovation, M. Breque, L. De Nul, A. Petridis, Industry 5.0 – Towards a sustainable, human-centric and resilient European industry, Publications Office of the European Union, 2021. doi:10.2777/308407.

[2] A. Sarkar, Enough with "human-ai collaboration", in: Extended Abstracts of the 2023 CHI Conf. on Human Factors in Computing Systems, 2023, pp. 1–8.

[3] J. Bütepage, D. Kragic, Human-robot collaboration: From psychology to social robotics, ArXiv abs/1705.10146 (2017).

[4] B. Sarkar, A. Shih, D. Sadigh, Diverse conventions for human-ai collaboration, Advances in Neural Information Processing Systems 36 (2024).

[5] S. Daronnat, L. Azzopardi, M. Halvey, Impact of agents' errors on performance, reliance and trust in human-agent collaboration, in: Proc. of the Human Factors and Ergonomics Society Annual Meeting, volume 64, SAGE Publications Sage CA: Los Angeles, CA, 2020, pp. 405–409.

[6] A. Borboni, K. V. V. Reddy, I. Elamvazuthi, M. S. AL-Quraishi, E. Natarajan, S. S. Azhar Ali, The expanding role of ai in collaborative robots for industrial applications: a systematic review of recent works, Machines 11 (2023) 111.

[7] N. Sebanz, H. Bekkering, G. Knoblich, Joint action: bodies and minds moving together, Trends in cognitive sciences 10 (2006) 70–76.

[8] E. M. Van Zoelen, K. Van Den Bosch, M. Neerincx, Becoming team members: Identifying interaction patterns of mutual adaptation for human-robot co-learning, Frontiers in Robotics and AI 8 (2021).

[9] K. van den Bosch, T. Schoonderwoerd, R. Blankendaal, M. Neerincx, Six challenges for human-ai co-learning, in: Adaptive Instructional Systems: 1st Int. Conf., AIS 2019, Held as Part of the 21st HCI Int. Conf., HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proc. 21, Springer, 2019, pp. 572–589.

[10] S. Holter, M. El-Assady, Deconstructing human-ai collaboration: Agency, interaction, and adaptation, arXiv preprint arXiv:2404.12056 (2024).

[11] M. Vössing, N. Kühl, M. Lind, G. Satzger, Designing transparency for effective human-ai collaboration, Information Systems Frontiers 24 (2022) 877–895.

[12] A. Shafti, J. Tjomsland, W. Dudley, A. A. Faisal, Real-world human-robot collaborative reinforcement learning*, IEEE/RSJ Int. Conf. on Intel. Robots and Systems (IROS) (2020).

[13] A. C. Tsitos, M. Dagioglou, Enhancing team performance with transfer-learning during real-world human-robot collaboration (2022).

[14] J. W. Krakauer, A. M. Hadjiosif, J. Xu, A. L. Wong, A. M. Haith, Motor learning, Compr Physiol 9 (2019) 613–663.

[15] P. Spitzer, N. Kühl, M. Goutier, Training novices: The role of human-ai collaboration and knowledge transfer, arXiv preprint arXiv:2207.00497 (2022).

[16] A. K. Dhawale, M. A. Smith, B. P. Ölveczky, The role of variability in motor learning, Annual review of neuroscience 40 (2017) 479–498.

[17] Z. Zhu, K. Lin, J. Zhou, Transfer learning in deep reinforcement learning: A survey, arXiv preprint arXiv:2009.07888 (2020).

[18] D. Honerkamp, T. Welschehold, A. Valada, Learning kinematic feasibility for mobile manipulation through deep rl, IEEE Robotics and Automation Letters 6 (2021) 6289–6296.

[19] A. Malik, Y. Lischuk, T. Henderson, R. Prazenica, A deep rl approach for inverse kinematics solution of a high degree of freedom robotic manipulator, Robotics 11 (2022).

[20] M. Q. Mohammed, K. L. Chung, C. S. Chyi, Review of deep reinforcement learning-based object grasping: Techniques, open challenges, and recommendations, IEEE Access 8 (2020).

[21] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Mueller, V. Koltun, D. Scaramuzza, Champion-level drone racing using deep reinforcement learning, Nature 620 (2023) 982–987.

[22] T. Haarnoja, A. Zhou, S. Ha, J. Tan, G. Tucker, S. Levine, Learning to walk via deep reinforcement learning, ArXiv (2018).

[23] F. Lygerakis, M. Dagioglou, V. Karkaletsis, Accelerating human-agent collaborative reinforcement learning, in: Proc. of the 14th PErvasive Technologies Related to Assistive Environments Conf., 2021, pp. 90–92.

[24] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of go with deep neural networks and tree search, Nature 529 (2016).

[25] N. Brown, T. Sandholm, Superhuman ai for multiplayer poker, Science 365 (2019).

[26] M. F. A. R. D. T. (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, A. P. Jacob, M. Komeili, K. Konath, M. Kwon, A. Lerer, M. Lewis, A. H. Miller, S. Mitts, A. Renduchintala, S. Roller, D. Rowe, W. Shi, J. Spisak, A. Wei, D. Wu, H. Zhang, M. Zijlstra, Human-level play in the game of diplomacy by combining language models with strategic reasoning, Science 378 (2022) 1067–1074.

[27] M. Carroll, R. Shah, M. K. Ho, T. L. Griffiths, S. A. Seshia, P. Abbeel, A. Dragan, On the utility of learning about humans for human-AI coordination, 2019.

[28] Y. C. Huang, Y. T. Cheng, L. L. Chen, J. Y. J. Hsu, Human-ai co-learning for data-driven ai, ArXiv (2019).

[29] H. Nguyen, H. La, Review of deep reinforcement learning for robot manipulation, in: 3rd IEEE Int. Conf. on Robotic Computing (IRC), 2019, pp. 590–595.

[30] Z. Zhu, K. Lin, A. K. Jain, J. Zhou, Transfer learning in deep rl: A survey, IEEE Trans. on

Pattern Analysis and Machine Intelligence 45 (2023).

[31] M. Islam, The impact of transfer learning on ai performance across domains, Journal of AI General science (JAIGS) 1 (2024).

[32] A. Ng, D. Harada, S. J. Russell, Policy invariance under reward transformations: Theory and application to reward shaping, in: Int. Conf. on Machine Learning, 1999.

[33] M. Vecerík, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. M. O. Heess, T. Rothörl, T. Lampe, M. A. Riedmiller, Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards, ArXiv (2017).

[34] S. Schaal, Learning from demonstration, in: Proc. of the 9th Int. Conf. on Neural Information Processing Systems, 1996, p. 1040–1046.

[35] M. Yang, O. Nachum, Representation matters: Offline pretraining for sequential decision making, in: Int. Conf. on Machine Learning, 2021.

[36] T. Hester, M. Vecerík, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, G. Dulac-Arnold, J. P. Agapiou, J. Z. Leibo, A. Gruslys, Deep q-learning from demonstrations, in: AAAI Conf. on AI, 2017.

[37] G. E. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, ArXiv (2015).

[38] F. Fernández, M. Veloso, Probabilistic policy reuse in a reinforcement learning agent, in: Proc. of the 5th Int. Joint Conf. on Autonomous Agents and Multiagent Systems, 2006.

[39] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep reinforcement learning that matters, in: Proc. of the 32nd AAAI Conf. on AI and 13th Innovative Applications of AI Conf. and 8th AAAI Symposium on Educational Advances in AI, 2018.

[40] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, A. Madry, Implementation matters in deep rl: A case study on ppo and trpo, in: Int. Conf. on Learning Representations, 2020.

[41] T. Eimer, M. Lindauer, R. Raileanu, Hyperparameters in reinforcement learning and how to tune them, in: Proc. of the 40th Int. Conf. on Machine Learning, 2023.

[42] K. van den Bosch, T. Schoonderwoerd, R. Blankendaal, M. Neerincx, Six challenges for human-ai co-learning (2019).

[43] P. Chattopadhyay, D. Yadav, V. Prabhu, A. Chandrasekaran, A. Das, S. Lee, D. Batra, D. Parikh, Evaluating visual conversational agents via cooperative human-ai games, in: AAAI Conf. on Human Computation & Crowdsourcing, 2017.

[44] G. Hoffman, Evaluating fluency in human–robot collaboration, IEEE Transactions on Human-Machine Systems 49 (2019) 209–218.

[45] D. Koutrintzes, Knowledge transfer in human-artificial intelligence collaboration, Master's thesis, University of Piraeus, 2023.

[46] P. Christodoulou, Soft actor-critic for discrete action settings, arXiv preprint arXiv:1910.07207 (2019).

[47] Y. Xu, D. Hu, L. Liang, S. McAleer, P. Abbeel, R. Fox, Target entropy annealing for discrete soft actor-critic, 2021.

[48] R. Liu, J. Y. Zou, The effects of memory replay in reinforcement learning, 2018 56th Annual Allerton Conf. on Communication, Control, and Computing (Allerton) (2017) 478–485.