Perceiving Agent Collaborative Sonic Exploration In Interactive Reinforcement Learning

Hugo Scurto, Frédéric Bevilacqua, Baptiste Caramiaux

PERCEIVING AGENT COLLABORATIVE SONIC EXPLORATION IN INTERACTIVE REINFORCEMENT LEARNING

Hugo Scurto
Ircam - Centre Pompidou
STMS IRCAM–CNRS–SU
Hugo.Scurto@ircam.fr

Frédéric Bevilacqua
Ircam - Centre Pompidou
STMS IRCAM–CNRS–SU
Frederic.Bevilacqua@ircam.fr

Baptiste Caramiaux
CNRS–LRI, Université Paris-Sud
STMS IRCAM–CNRS–SU
Baptiste.Caramiaux@lri.fr

ABSTRACT

We present the first implementation of a new framework for sound and music computing, which allows humans to explore musical environments by communicating feedback to an artificial agent. It is based on an interactive reinforcement learning workflow, which enables agents to incrementally learn how to act on an environment by balancing exploitation of human feedback knowledge and exploration of new musical content. In a controlled experiment, participants successfully interacted with these agents to reach a sonic goal in two cases of different complexities. Subjective evaluations suggest that the exploration path taken by agents, rather than the fact of reaching a goal, may be critical to how agents are perceived as collaborative. We discuss such quantitative and qualitative results and identify future research directions toward deploying our "co-exploration" approach in real-world contexts.

1. INTRODUCTION

When creating music, musicians make use of various forms of exploration to achieve their musical goals. Such exploration is essential to facilitate expression and discovery along their creative process [1]. In the specific case of music computing, exploration generally consist in probing some parameter space—for example to grasp a synthesizer's sonic abilities.

To facilitate exploration, system designers must find a compromise between the system's complexity and its musical expressiveness. This is often a difficult design constraint. As high expressiveness tends toward high complexity—and reversely, low complexity tends toward low expressiveness [2]—, music computing systems often result in complex interfaces which require expert knowledge to be explored. The standard VST constitutes the typical example of such a design approach, usually featuring tens of ad hoc knobs and thousands of potential combinations.

In the last decade, interactive supervised learning approaches have enabled to overcome this difficulty, by combining powerful computational abilities (such as autonomous learning, recognition or prediction) with human-centred interaction (such as generating new content from direct user examples) [3]. Several systems have since been designed, with successful applications in music composition, performance, and pedagogy [4–7]. Interestingly, many musicians reported that these interactive systems offered space for exploration [8], often personifying them as *collaborative partners* because of their ability to learn implicit musical properties similarly to a human musical collaborator [3, 9].

We are interested in designing a framework enabling exploration through interactive machine learning, in order to reinforce this sense of collaborative partnership. For this, we propose to provide musicians with the possibility to explore musical environments by communicating feedback to the system. Such an approach could be useful in cases where giving high-level feedback would be preferred over parameterizing a given system at a low-level. For example, users could progressively shape musical contents and processes using subjective evaluations.

In this paper, we describe first steps toward such a feedback-based interactive learning system for exploration. We propose to investigate interactive reinforcement learning as a new paradigm to interaction with musical environments. Reinforcement learning algorithms (also called *agents*) differ from supervised learning algorithms in the sense that they are capable of learning incrementally from acting on their environment. In the prototype that we implemented, any parameter space can be jointly explored by an agent (that directly acts at a parameter level) and a human (that gives positive or negative feedback to the agent regarding its current action). By iteratively giving feedback to the agent, users should be able to progressively shape the parameter space according to their subjective evaluations, thus potentially paving the way for collaboration.

We have led a controlled experiment with human participants interacting with such agents. Our approach differs from other reinforcement learning works in the sense that we aim at studying the exploration path taken by agents in an interactive setup, rather than their ability to learn a good behaviour. We found that agents are able to reach a goal in parameter spaces of varying complexities by balancing exploitation of human feedback knowledge and exploration of new musical content. Quantitative analysis of subjective evaluations suggest that the path taken by agents, rather than the fact of reaching a goal, may be critical to perceiving collaboration during exploration. These results provide a baseline understanding for future implementations and real-world investigations of interactive reinforcement learning for collaborative exploration.

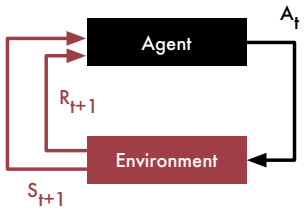

Figure 1. A standard reinforcement learning, where an agent learns from its environments by directly acting on it.

2. BACKGROUND AND RELATED WORK

Machine learning algorithms have long been used in the context of sound and music computing. In the following section, we review research where user interaction is central to the design of such algorithms.

2.1 Standard Interactive Learning Setups

2.1.1 Interactive Supervised Learning

Supervised learning have been investigated for interactively designing motion-sound mappings [7,10]. User interaction with supervised learning consists in demonstrating example gestures to the learning algorithm so that it can learn to recognize them on the fly. Exploration within supervised learning either takes place during the training phase (where users can experiment with several examples of different gestures), or during the running phase (where they can explore interpolations between given examples) [8]. This two-phase workflow has been shown useful for a number of tasks; however, it has been shown limiting in some cases, for example when users want to slightly modify a given design [11]. Also, it does not support sequential adaptation to how users explore the system, which may be critical for our use case.

2.1.2 Online Learning Agents

Sequential adaptation have been investigated for interacting with autonomous agents [12]. User interaction with autonomous agents consists in generating example musical content for guiding agents' musical behavior. Exploration within autonomous agents mainly consists in continuous musical improvization with the agent (through sound [13] or motion [14]). This online learning workflow has been shown useful for performance cases (which require continuous generation and reactivity) [9,15] but may not be fully adapted to more general, "offline" design cases. Crucially, it still coerces users into demonstrating musical examples to explore new behaviors, which might be limiting in cases where users may not have an initial idea in mind [11], and which might prevent users from communicating more general, high-level feedback to the system on how it should behave.

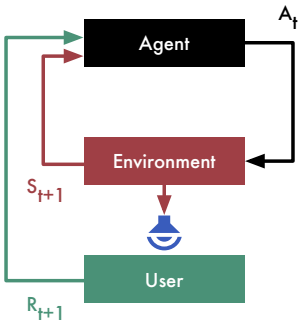

Figure 2. Our interactive reinforcement learning workflow, where the user is responsible for giving a reward to the agent as a consequence of its actions.

2.2 Other Interactive Learning Setups

2.2.1 "Creative" Machine Learning

Recently, new interactions with supervised learning algorithms have been explored in the context of motion-sound mapping design. Scurto et al. implemented a machine learning tool able to generate many alternative user-adapted mappings from only one motion stream [11]. This workflow avoided users to reflect on what examples they should demonstrate for reaching a goal: rather, it enabled them to focus only on subjective, evaluative exploration of many prototypes. Users valued the space for exploration offered by such autonomous generation abilities. However, they expressed a lack of control over the system, as generation remained fully autonomous and not adaptive—neither sequentially, nor subjectively.

2.2.2 Interactive Reinforcement Learning

There is still relatively few works investigating the *interactive* uses of reinforcement learning in the field of music computing [3]. Derbinsky et al. [16] proposed to use reinforcement learning to model rhythms played by human performers, but do not integrate user interaction so as to learn subjective evaluations.

In parallel, research in other fields such as robotics [17] and computer science [18, 19] have made huge progress toward the development of interactive agents capable of learning from human feedback. These agents support sequential adaptation without needing example demonstrations, but only by receiving human feedback as subjective evaluations of the autonomously-generated behaviors. We believe such interactive reinforcement learning workflow should enable the creation of novel collaborative partners in musical exploration, and foster new creative applications in sound and music computing.

3. SYSTEM AND WORKFLOW DESCRIPTION

In this section, we describe our interactive reinforcement learning framework, from the standard reinforcement learning algorithms (see Figure 1) to its implementation adding user interaction (see Figure 2 and 3). We finally describe the prototype that we implemented for our experiment.

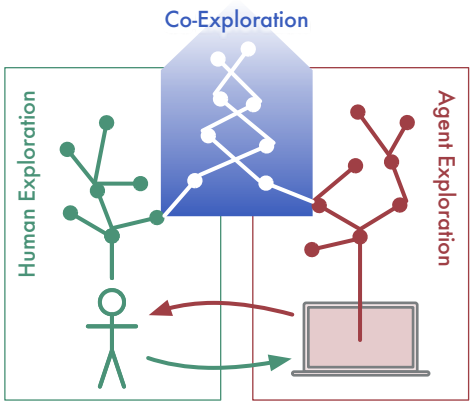

Figure 3. Our "co-exploration" workflow, standing for collaborative human exploration of musical content and agent exploration of parameter space.

3.1 System and Workflow

3.1.1 Reinforcement Learning and Exploration

Reinforcement learning defines a family of algorithms—called "agents"—able to learn from interacting with their environment (see Figure 1) [20]. More formally, at each time step t, the agent receives some representation of the environment's state, $S_t \in \mathcal{S}$, and on that basis selects an action $A_t \in \mathcal{A}(S_t)$. One time step later, in part as a consequence of its action, the agent receives a numerical reward from the environment, R_{t+1}, and finds itself in a new state, S_{t+1}. At each time step, based on the reward it received, the agent updates a policy that maps each state to probabilities of selecting each possible action. This update is computed so as to maximize the total amount of reward expected to be received over the long run.

Importantly, reinforcement learning agents must keep on trying new actions to discover which ones yield the most reward. In this kind of situation, the agent is said to be *exploring* the environment. When the agent takes the action that maximizes future reward from a given state, the agent is said to *exploit* what it has learned. By taking such exploratory path—balancing between exploitation and exploration—, reinforcement learning agents are able to progressively learn an optimal policy toward their environment, and converge to an optimal interactive behavior. Several methods for environment exploration and policy updating have been studied, and constitute the core of current reinforcement learning research [20].

3.1.2 User Interaction and Co-Exploration

To implement interaction with such agents, we must modify the formal framework defined above. We propose, along with previous works [17–19], that a human would be responsible for giving reward to the agent (see Figure 2). Our hypotheses are that the numerical reward may constitute a feedback channel from the human to the agent (respectively giving positive, zero, or negative reward for positive, neutral, or negative feedback), and that interactively communicating feedback toward the environment following the agent's exploration path may support human exploration.

Importantly, as previously said, we would like to provide musicians with tools to explore digital environments. In our implementation, users would be allowed to do so by giving feedback to the agent—at a subjective, high level. Parallel to users, the agent would also be exploring the musical environment—at a parameter, low level. Therefore, we propose to call "co-exploration" such an approach of collaborative human exploration of musical content and agent exploration of parameter space (see Figure 3). This is a non-trivial problem, and our aim is to provide a first understanding of the challenges at stake through our case study and discussion (in Sections 4 and 5).

3.2 Implementation

3.2.1 Current Prototype

We are currently implementing coax, a Python software library for collaborative human-agent exploration. It allows to connect agents to any kind of musical environment that sends and receives OSC messages. The current prototype implements the Sarsa learning algorithm [20]. It is a state-action, one-step, model-free, tabular method, which respectively means that (1) it learns which actions to take in a given state, (2) it updates the agent's policy at each time step, (3) it learns for state-actions that it has visited, and (4) it works for discrete, finite state-action spaces. Action selection is performed using the ε-greedy method [20], which is selecting a random action (*exploration*) with probability ε, and selecting the action that has highest value (*exploitation*) with probability $1 - \varepsilon$. We will discuss further implementation improvements in Section 5.

3.2.2 Use Case Formalization

The following use case focuses on one common type of digital musical environment: VST instruments. We propose the following formalization in the frame of interactive reinforcement learning. The environment's state consists of a vector of VST parameters; the agent's actions consists of moving one of these parameters up or down (except for VST boundary values, that the agent cannot exceed). At each time step, the agent generates a sound by acting on the VST: then, the human communicates feedback to the agent regarding its subjective evaluation of the generated sound. The more the agent receives feedback information, the more it should converge to the human's goal. Many other formalizations could be considered—we will discuss them in Section 5.

4. CASE STUDY: EVALUATING HUMAN PERCEPTION OF AGENTS

As a first step toward co-exploration, we led a controlled experiment with human participants. Our aim is to study how the path taken by agents during exploration may be perceived and influenced by humans in an interactive setup.

4.1 Method

4.1.1 Participants

We recruited 12 participants (average of 26.9 years old, $\sigma = 7.44$, 5 Female and 7 Male). Half of them were music computing practitioners. All of them reported normal hearing.

4.1.2 Task

The basic task of the study was to guide an agent through a VST, from the lowest to the brightest sound. At each step of the task, the agent would generate a new sound. If the new sound was brighter than the previously generated one, participants had to give positive feedback to the agent. In any other cases (lower or similar brightness), participants had to give negative feedback to the agent. The task automatically ended in two cases: either the brightest sound was reached, or it was not reached after a maximum number of steps (we set it to 150).

At the end of the task, participants were asked to rate their perception of the agent according to three aspects related to collaboration. The first aspect was the degree of agency provided by the agent through feedback (*"did the agent seem to take into account your feedback in a reactive manner, or did it seem to act completely independently?"*). The second aspect was the degree of assistance provided by the agent throughout the task (*"did the agent seem to generate sounds that were brighter, or did it seem not to be of any help in going in this direction?"*). The third and last aspect was the degree of easiness of the task (*"overall, did the task seem to be very easy, or very difficult?"*).

4.1.3 Agents

Three types of agents were evaluated: "random", "balance", and "exploit". These correspond to three different degrees of exploration ($\varepsilon = 0$: the agent only takes random actions; $\varepsilon = 0.5$: the agent balances random action selection with feedback-based best action selection with probability 0.5; $\varepsilon = 1$: the agent only selects the best actions as indicated by user feedback). Other agent parameters were fixed so that exploration would be the sole varying factor.

4.1.4 Musical Environments

Sounds were generated through a FM synthesis engine (implemented in Max/MSP), with two discretized parameters. The first parameter, called modulation index, could take ten values ranging from 3 to 70; the second parameter, called harmonicity ratio, could take three values ranging from 0.98 to 1.02. The resulting VST thus had 30 possible states, corresponding to 30 static sounds. As previously explained, the agent's possible actions consist in moving up or down one of the two parameters. For the sake of the experiment, we normalized sound loudness empirically so they perceptually appear of equal intensity, and we set sound duration to 500 ms.

Based on this VST, we designed two environment models in close relationship with the task's goal: "unobstrucked", and "obstrucked" (see Figure 4). In the unobstrucked environment, brightness increases linearly with modulation index: highest brightness thus corresponds to highest index value. We expect "balance" and "exploit" agents to be more collaborative than "random" agents through their ability to learn and select the best actions.

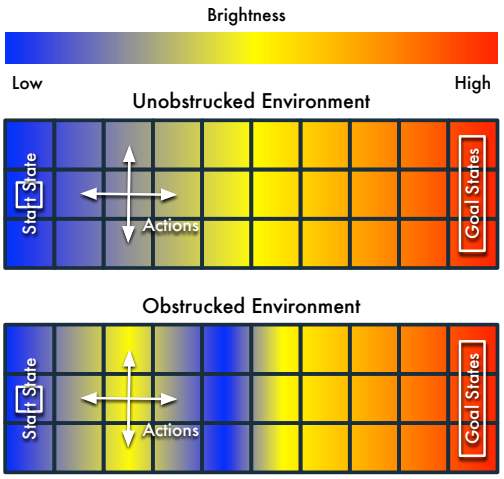

Figure 4. The two environment models designed for our experiment. Top: Unobstrucked environment, where brightness varies linearly. Bottom: Obstrucked environment, where brightness varies nonlinearly.

In the obstrucked environment, brightness varies nonlinearly with modulation index: highest brightness still corresponds to highest index value, but a local maximum lives at one third of the scale. Our hypothesis is that "exploit" agents would remain stuck in this local maximum, whereas "balance" agents would overcome it through their ability to explore. We thus expect "balance" agents to be more collaborative than "random" and "exploit" agents.

4.1.5 Procedure

The experimental session consisted of a familiarization phase and an experimental phase.

Participants first had to read the task's instruction and could ask the experimenter for clarification if necessary. Then, they had two test tasks in the unobstrucked environment with two types of agents (one "exploit", then one "random") to familiarize with the range of sounds and agent behaviors at stake. Sounds were presented as pairs to participants (using headphones), so as to facilitate brightness comparison between the previously-generated sound and the new one. Participants could listen to a pair of sounds as many time as they wanted to (using a keyboard key) before giving positive or negative feedback to the agent (using left or right arrow keys). Once a task was over, participants had to rate the agent's behavior for each of the three previously-described aspects on 9-point Likert scales (using the mouse and interactive sliders). We asked participants to use the full scales as much as they could.

Once this phase was over, participants could start the experimental phase. The first stage only concerned the unobstrucked environment: participants were asked to guide and evaluate each of the three types of agents within it. For improving consistency, participants made three trials with each of the three agents. A stage thus consisted in nine

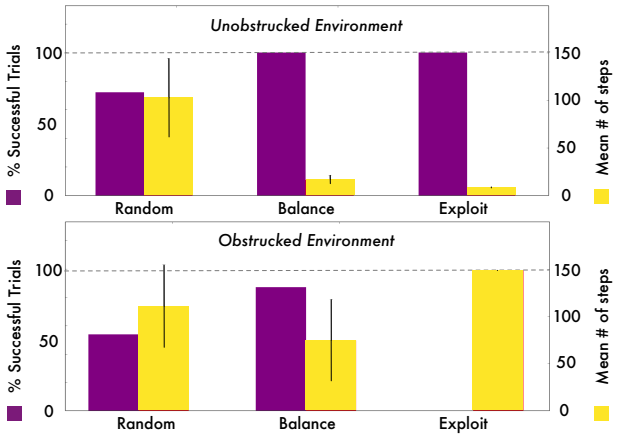

Figure 5. Synthetic trial data.

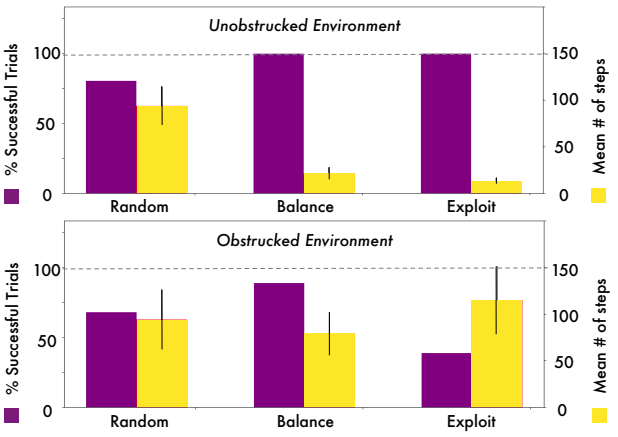

Figure 6. Participants' trial data.

tasks that were randomized in order. Finally, the second stage only concerned the obstructed environment: similarly, participants guided and evaluated the three types of agents three times each, in a random order. Participants were allowed to take a break at any time during the session, which lasted one hour on average.

4.2 Results

For each participant, we recorded step-by-step data (time, states, actions, feedback and ratings), as well as audiovisual data of users. Prior to analysing them, we report on synthetic data generated before the actual experiment.

4.2.1 Synthetic Trial Data

We programmed synthetic feedback users of same number as participants to generate a benchmark on how agents should ideally behave in our two environment models. This case corresponds to participants giving perfectly consistent feedback.

We measured the percentage of successful trials (which reflects the probability of reaching the goal), as well as the mean number of steps taken in a trial (which reflects a trial's duration), for each type of agent and in each of the two environments (see Figure 5). For each environment, we submitted each measure to a one-way ANOVA with agent exploration as the within-subject factor. In the unobstructed environment, the effect of exploration was significant for both percentage of successful trials $[F(2, 22) = 8.83, p < 0.001]$ and mean number of steps $[F(2, 22) = 91.3, p < 0.001]$. Planned contrasts showed that both measures significantly differed for "balance" and "exploit" agents compared to "random" agents.

Likewise, in the obstructed environment, the effect of exploration was significant for number of successful trials $[F(2, 22) = 44.7, p < 0.001]$ and mean number of steps $[F(2, 22) = 26.3, p < 0.001]$. Planned contrasts showed that both measures significantly differed for "balance" agents compared to "random" and "exploit" agents.

4.2.2 Participants' Trial Data

We first measured participants' feedback behavior. In the unobstructed environment, participants gave a mean of 393 feedback every 1.91 s, with 96.3% being correct. In the obstructed environment, participants gave a mean of 879 feedback every 1.84 s, with 98.0% being correct.

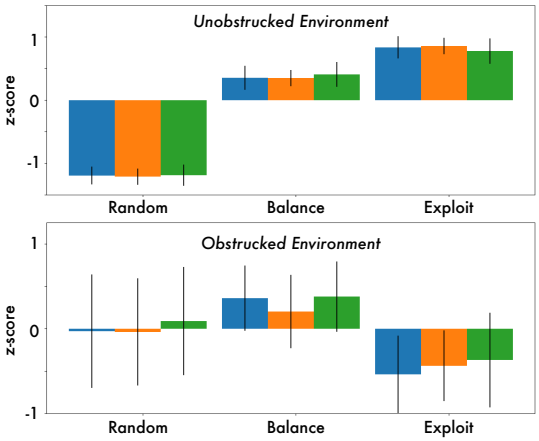

Figure 7. Participants' evaluation data. In blue: agency. In orange: assistance. In green: easiness.

Similarly to synthetic users, we measured the percentage of successful trials, as well as the mean number of steps taken by each of the three agent types, in each of the two environments (see Figure 6). We used the mean of all trials in each condition for each participant. For both environments, we submitted both measures to a one-way ANOVA with agent exploration as the within-subject factor. In the unobstructed environment, the effect of exploration was significant for percentage of successful trials $[F(2, 22) = 6.49, p < 0.005]$ and mean number of steps $[F(2, 22) = 130.3, p < 0.001]$. Planned contrasts showed that both measures significantly differed for "balance" and "exploit" agents compared to "random" agents.

Likewise, in the obstructed environment, the effect of exploration was significant for percentage of successful trials $[F(2, 22) = 8.16, p < 0.002]$ and mean number of steps $[F(2, 22) = 3.62, p < 0.03]$. Planned contrasts showed that both measures significantly differed for "balance" agents compared to "random" and "exploit" agents.

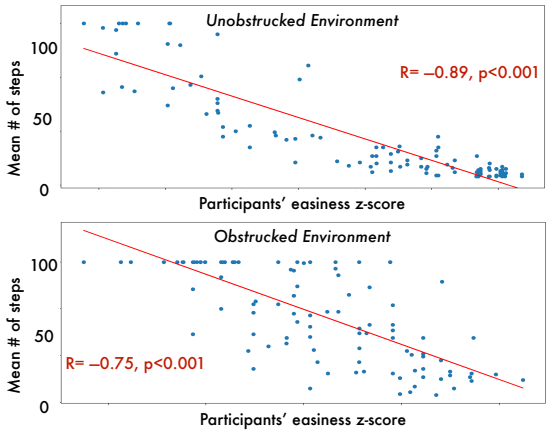

Figure 8. Participants' ratings versus task parameters.

4.2.3 *Participants' Evaluation Data*

We computed the standard score (also called z-score) for each evaluation ratings in each environment to compare participants on the same scale (see Figure 7).

For each environment, we submitted each z-score to a one-way ANOVA with agent exploration as the within-subject factor. In the unobstrucked environment, the effect of exploration was significant for all three perceptual aspects ($[F(2, 22) = 429.3, p < 0.001]$ for agency; $[F(2, 22) = 767.3, p < 0.001]$ for assistance; and $[F(2, 22) = 335.2, p < 0.001]$ for easiness). Planned contrasts showed that all three perceptual ratings were significantly higher for "balance" and "exploit" agents than for "random" agents.

Likewise, in the obstrucked environment, the effect of exploration was significant for for all three perceptual aspects ($[F(2, 22) = 8.32, p < 0.002]$ for agency; $[F(2, 22) = 4.53, p < 0.02]$ for assistance; and $[F(2, 22) = 5.26, p < 0.02]$ for easiness). Planned contrasts showed that all three perceptual ratings were significantly higher for "balance" agents than for "random" and "exploit" agents.

Finally and as shown in Figure 8, we measured that participants' perception of task easiness was correlated with the total number of steps taken by all types of agents, in both environments.

5. DISCUSSION AND FUTURE DEVELOPMENTS

In this section, we discuss our experiment's results and extract implications for our system's future developments.

5.1 The usefulness of balancing exploitation with exploration

5.1.1 *Synthetic trial data*

We first look at synthetic trial data to analyse agents' ability to reach a goal in a non-interactive setup. In the unobstrucked environment, as expected, all agents that took into account feedback ("balance" and "exploit") always succeeded in reaching the goal, with "exploit" agents being the fastests as they took the best action at each step; "random" agents reported the worst performance, succeeding only two thirds of the time with lower speed. In the obstrucked environment, conversely, "exploit" agents never succeeded in reaching the goal. As expected, they remained stuck in the local maximum that we designed. In this case where an obstacle blocks the way to the goal, "balance" agents remarkably outperformed other agents in both speed and number of success. This proves that agents' balance between exploitation and exploration may be useful for reaching a goal in environments of varying complexities.

5.1.2 *Participants' trial data*

Participants' trial data differ from synthetic trial data because of imperfect feedback occasionally given by users. Despite this difference, agents took exploration paths that were similar to those generated with synthetic users in five out of six agent-environment cases, as shown in Figure 6. In the remaining case of "exploit" agents exploring the obstrucked environment, one third of the trials were successful, which means that agents unexpectedly managed to overcome the obstacle that we designed to reach the goal. This proves that agents can take different paths in an interactive setup where users make feedback mistakes.

5.2 The influence of exploration path on user perception

5.2.1 *Perceiving collaboration*

We now analyse participants' subjective evaluations to better understand how exploration might be perceived by users. First, we observe that participants' ratings had more variability in the obstrucked environment than in the unobstrucked environment. This suggests that an environment's complexity may strongly influence how humans perceive agent exploration. Second, we noticed that participants rated down "exploit" agents in the obstrucked environment, even if one third of them succeeded in reaching the goal, as we previously discussed. This proves that the path taken by agents during exploration may be more critical to how collaborative agents are perceived by users than the actual fact of reaching the goal.

Looking more in detail to participants' ratings, we can see that "balance" agents were the only type of agents that were perceived as being the most assistive in both environments, thus reflecting their quantitative usefulness. As expected, "random" agents were perceived as providing the less agency in the unobstrucked environment: this suggests that participants may be able to perceive when an agent learns along its path—in other words, there was no "placebo effect" toward agents' artificial intelligence. Finally, even if "exploit" agents formally take the best action at every step as defined by participants' feedback, this may not be perceived by participants, as their ratings of agency shows (see Figure 7, bottom). This confirms that an agent's internal functioning may not be properly perceived by humans, whose perception might be more influenced by the path taken by agents in a given environment. Results shown on Figure 8 seem to confirm this statement, as one of the evaluation ratings correlates with one of the task parameters, regardless of the type of agent at stake.

5.2.2 Personifying agents

Interestingly, audiovisual recordings shows that all participants personified agents depending on their perceived collaboration. For example, agents that took relatively direct paths to the goal provoked positive reactions (such as *"it understood right away"*) and adjectives (e.g., *"nice"*, or *"careful"*). On the other hand, agents that took more complex paths—such as "random" agents, or "exploit" agents that remained stuck in the obstacle—inherited depreciative reactions (e.g., *"it doesn't listen to me"*, or *"it seems light-headed"*) and adjectives (e.g., *"idiot"*). This might be a first clue—to some extent—for stating that feedback-based interaction may encourage users to perceive agents as human-like partners—in some cases able to act as collaborators.

5.3 Towards co-exploration

5.3.1 The issue of human moving goals

In our experiment, we forced participants to follow a fixed feedback strategy: this might limit the reach of our experiment's results. Indeed, such feedback constraint might not be realistic in real-world exploration, mainly for two points: (1) users might change their feedback strategy, and (2) their goals might evolve over time. These situations are typical of real-world scenarios, where users may push agents in limit conditions [2], or may want to explore several alternative strategies [8]. Investigating these points constitute next steps toward turning our interactive reinforcement learning system (where the goal to be learned was fixed) into a co-exploration system (where the goal to be learned might evolve as the human uses the system).

5.3.2 Improving algorithms or interactions?

We identify two main directions for addressing these points—stressing that these directions could be complementary. The first option corresponds to investigate other reinforcement learning algorithms. As said, our current prototype implements the Sarsa algorithm, which is a standard method for reinforcement learning. Other approaches to learning may be better adapted to our co-exploration use case. For example, one may investigate methods that are robust to non-stationary feedback [18]. Alternatively, one may also investigate approximate policy learning algorithms [19,20] for learning relevant representations of an environment without having to explore it in its entirety.

The second option corresponds to design new interactions that may better fit interactive uses of reinforcement learning algorithms. As shown in our study, humans may not always perceive how a learning system internally works. In order to give more control to the human, one could imagine allowing humans to modify agent parameters during interaction, for example by actively choosing the degree of exploration they may need. Also, one could allow humans to go backwards in the agent's learning process, or to restart learning at any time, so as to give space for iterative, flexible exploration patterns [1]. Again, all these developments are not contradictory, and we believe that both directions should be considered in future research.

5.3.3 Connecting agents to real-world systems and situations

Finally, our experiment focused on models of musical environments whose dimensionalities may not fully reflect those of standard music computing systems to be explored by users. Yet, we argue that investigating such models have provided useful insights on how agents would take exploration paths in real-world music systems. We are currently leading several studies connecting our current system with other VSTs as well as motion-sound mapping models, hoping to harvest complementary insights on our use case and pushing further the formalization of environments at stake in our co-exploration agents.

Such studies might also be an opportunity for investigating other qualitative methods for evaluating agents. Indeed, our experiment's results suggested that participants did not really differentiate each of the three perceptual aspects they had to rate, which in turn suggest that they may have a much global appreciation of how an agent interact with them. Borrowing approaches and methods from the field of Human-Computer Interaction (such as user-centered design through case studies and workshops) [1,5] might be essential for grasping such experiential aspects among humans and for leading such situated studies with agents.

6. CONCLUSION

This paper presented first steps toward investigating interactive reinforcement learning agents for collaborative musical exploration. Its first contribution is a new interactive learning framework focusing on human exploration, where users would be allowed to guide an agent's learning process by communicating subjective feedback. A working implementation allowed the running of an experiment led with human participants, which constitutes the second contribution of this work. Results suggest that interactive reinforcement learning agents may be able to reach a goal by balancing exploitation of human feedback knowledge and exploration of new musical content, and that the path taken by agents during exploration may be critical to how collaborative agents are perceived by users.

Based on these results, we identified several directions to iterate the design of our system. Other reinforcement learning models may be investigated to provide agents with learning strategies better adapted to users. Other interactions with agents may be designed so as to put more control in the hands of humans. Finally, new musical case studies in real-world situations may be led to explore all such design alternatives. We believe iterating through these steps will enable to progressively converge to an optimal design of our interactive reinforcement learning framework, and may help better understand what makes an interactive learning agent a great musical co-explorer.

Acknowledgments

We thank all participants for their time and feedback. We also thank Benjamin Matuszewski and Bavo Van Kerrebroeck for useful discussion and suggestions.

7. REFERENCES

[1] M. Resnick, B. Myers, K. Nakakoji, B. Shneiderman, R. Pausch, T. Selker, and M. Eisenberg, "Design principles for tools to support creative thinking," *Working Paper*, 2005.

[2] S. Jorda, "Digital lutherie crafting musical computers for new musics' performance and improvisation," Ph.D. dissertation, Universitat Pompeu Fabra, 2005.

[3] R. Fiebrink and B. Caramiaux, "The machine learning algorithm as creative musical tool," *arXiv preprint arXiv:1611.00379*, 2016.

[4] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana, "Continuous realtime gesture following and recognition," in *International gesture workshop*. Springer, 2009, pp. 73–84.

[5] R. Fiebrink, P. R. Cook, and D. Trueman, "Human model evaluation in interactive supervised learning," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 147–156. [Online]. Available: http://doi.acm.org/10.1145/1978942.1978965

[6] B. Caramiaux, N. Montecchio, A. Tanaka, and F. Bevilacqua, "Adaptive gesture recognition with variation estimation for interactive systems," *ACM Trans. Interact. Intell. Syst.*, vol. 4, no. 4, pp. 18:1–18:34, Dec. 2014. [Online]. Available: http://doi.acm.org/10.1145/2643204

[7] J. Françoise, N. Schnell, R. Borghesi, and F. Bevilacqua, "Probabilistic models for designing motion and sound relationships," in *Proceedings of the 2014 international conference on new interfaces for musical expression*, 2014, pp. 287–292.

[8] R. Fiebrink, D. Trueman, N. C. Britt, M. Nagai, K. Kaczmarek, M. Early, M. Daniel, A. Hege, and P. R. Cook, "Toward understanding human-computer interaction in composing the instrument." in *ICMC*, 2010.

[9] F. Pachet, "The continuator: Musical interaction with style," *Journal of New Music Research*, vol. 32, no. 3, pp. 333–341, 2003.

[10] B. Caramiaux and A. Tanaka, "Machine learning of musical gestures." in *NIME*, 2013, pp. 513–518.

[11] H. Scurto, R. Fiebrink *et al.*, "Grab-and-play mapping: Creative machine learning approaches for musical inclusion and exploration," in *Proceedings of the 2016 International Computer Music Conference*, 2016.

[12] G. Assayag, "Creative symbolic interaction," in *40th Intl. Comp. Mus. Conf. and 11th Sound and Music Comp. Conf.(ICMC/SMC joint conf.)*. ICMA, SMC, National and Kapodistrian University of Athens, IRMA, 2014, pp. pp–1.

[13] J. Nika, K. Déguernel, A. Chemla, E. Vincent, G. Assayag *et al.*, "Dyci2 agents: merging the" free"," reactive", and" scenario-based" music generation paradigms," in *International Computer Music Conference*, 2017.

[14] H. Scurto, F. Bevilacqua, and J. Françoise, "Shaping and exploring interactive motion-sound mappings using online clustering techniques," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME'17)*, 2017.

[15] G. Assayag, G. Bloch, M. Chemillier, A. Cont, and S. Dubnov, "Omax brothers: a dynamic yopology of agents for improvization learning," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. ACM, 2006, pp. 125–132.

[16] N. Derbinsky and G. Essl, "Exploring reinforcement learning for mobile percussive collaboration," in *Proceedings of the 2012 International Conference on New Interfaces for Musical Expression (NIME)*, 2012.

[17] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6-7, pp. 716–737, 2008.

[18] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The tamer framework," in *Proceedings of the fifth international conference on Knowledge capture*. ACM, 2009, pp. 9–16.

[19] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *arXiv preprint arXiv:1706.03741*, 2017.

[20] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press, 2011.