Ph.D proposal: Learning Human-like Bots in Video Games

Ludovic Denoyer and Sylvain Lamprier

January 31, 2023

1 Introduction and Objectives

Programming good bots in video games is still an open challenge. The usual way is to manually define a behavior tree that encodes the way the bot behaves together with the conditions that control when the agent has to switch from one behavior to another. This methodology has a major drawback: it is very time-consuming and needs a lot of effort to achieve a reasonable result. As a solution, (deep) reinforcement learning is a natural candidate to automatically discover efficient bots. This family of algorithms has recently achieved strong performance in games like starcraft, chess, go, etc. However, as a drawback, these methods are mostly designed with the objective to find optimal bots for a clearly specified task to achieve. For instance, in chess, reinforcement learning is used to learn a super-human play able to win all the games it plays. If it is a relevant approach to particular use cases, it is unsatisfactory from a gameplay perspective where the objective is to enhance fun of players, for which personalized play styles would be better suited. This also suffers from limitations for game analysis purposes.

The particular problem we face in this thesis is the problem of learning **bots that behave like humans**. Instead of looking for the best bot, our objective is to discover a population of bots that are able to capture the different play-styles of humans, e.g different play levels, different strategies, etc... Being able to learn human-like bots would have a very high impact for the video game industry, for different use-cases. A first obvious use-case is *player-facing bots* where such bots would play with/against humans, providing a realistic experience, or, even better, a personalized game experience to the player. A second but critical other use-case is the use of such bots to test games during production. Indeed, testing games is a major challenge that is currently done with artificial methods that are far from testing the systems in a realistic way. Being able to introduce bots that behave like humans would be a breakthrough, allowing to greatly reduce the cost of the conception of games by focusing the attention of the developers on what really matters to the players.

Using reinforcement learning for learning human-like bots is far from being a solved problem: first, RL techniques rely on a reward function that one wants to maximize. In this setting, the bot will be acting toward the achievement of a particular task specified through the reward function. As a consequence, if the reward is not aligned with the utility function used by the players, the resulting agent will solve the task in a completely unnatural way. Moreover, defining such a reward function may be impossible for general game playing since we don't know which utility function a human is optimizing, and different players are certainly optimizing different functions resulting in different play styles. There is no trivial way to adapt RL techniques to capture the behavior of humans.

The most promising direction to look at is the *Imitation Learning* and *Inverse Reinforcement Learning* settings [AD21] where one considers that we have access to a dataset of episodes (i.e data generated at each timestep of the execution of a game), this dataset being used to discover relevant behavior. Most of this literature has focused on the problem where such datasets contain expert traces (or at least a reward signal) with the idea to learn one unique policy that achieves the best possible performance. This is different to our setting where we want to learn multiple policies able to reproduce the diversity of human behaviors. There is thus a clear open research direction in finding ways to adapt these techniques to our particular problem (see Related Work Section)

The Ph.D thesis is oriented toward the goal of defining new techniques to learn bots that behave like humans. We propose to address this problem differently, by exploiting game traces captured when real players are playing the game. The Ph.D will be executed using the set of tools developed in La Forge which include realistic multiplayer 3D games made specifically for research, large datasets of traces and learning algorithms deployed on GPU clusters. More particularly we define three research axis, each of them mixing both theoretical and practical advances:

- **Axis 1: Capturing play-styles** In this first axis, the objective is to learn multiple bots that capture the diversity of human behaviors.
- **Axis 2: Replacing humans** As a second axis, the student will also focus on the problem of identifying the style of a particular player, to then be able to reproduce a particular playstyle, and, for instance, to replace the human player by the corresponding bot in the game.
- **Axis 3: Learning human-like controllable bots** As a last axis, we expect to extend our approach to allow to get more control on the style of each bot, for instance resulting in bots able to play like a particular player, but in a more offensive or defensive way. This axis will provide new design tools to create interesting games.

2 Research Program

2.1 Axis 1: Capturing play styles

As a first step, we expect to develop models that will learn not only one average behavior (as it is done when using classical imitation learning techniques), but multiple and diverse behaviors that capture the continuum of human play styles. It will be related to imitation learning with the objective to learn simultaneously imitation policies guided by a latent vector that captures the style of the player, the policy and the latent space being built together. Said otherwise, instead of learning a policy $\pi(a|s)$ optimal to a particular reward function, we expect to learn a style-conditioned policy $\pi(a|s,z)$, z being the style together with a distribution of style $p(z)$ such that it reproduce the diversity of human play styles. This axis relates to bayesian machine learning, employed for imitation learning with unsupervised behavior inference.

2.2 Axis 2: Replacing Humans

Being able to play like a human is already interesting for different applications like game testing. But it may be also interesting to have bots that can replace any player anytime in the game. For instance, if a player is disconnected, then such a bot would be able to continue to play the corresponding character in the same style as the disconnected player, avoiding changing the dynamics of the game. Such a system would need to play like a human (Axis 1), but also to be able to detect which playstyle a current player is adopting such that it will be possible to replace it anytime. If we denote $s_1, ..., s_t$ the state of the games when the player was playing, the objective is to continue the game by using the bot $\pi(a|s, f(s_1, ..., s_t))$ where $f(s_1, ..., s_t)$ is the function that is able to identify the style of the player from early interactions with the game. This axis relates to few-shot imitation learning, applied to games.

2.3 Axis 3: Learning human-like controllable bots

As a last step, we expect to augment our methods with the ability to build a meaningful controllable space. Indeed, if learning a latent space of styles allows one to define different styles through different vectors in this space, it cannot be used to understand what are the style dimensions that we control. It makes these approaches difficult to use by designers that want to define interesting behaviors from a gameplay perspective. To overcome this limit, we propose to enforce the model to map the latent space dimensions with meaningful information. As an example, one meaning full dimension is the play level of the bot which can be extracted directly from the dataset. Our objective will be to add a control space c on top of the play style, and to discover policies that satisfy both a particular play style z, but also the information specified in c. This axis relates to the disentanglement sub-field of machine learning, applied to complex behaviors.

3 Project Organization

Timeline. In the first 6 months, the student will study the RL literature and start to develop the first unsupervised RL approaches. In parallel, he will develop useful quantitative metrics and a benchmark to evaluate and test-bed CRL agents. Upon completion or partial results, the student will then move on to more ambitious projects in the 3rd , 4th and 5th semester. The 6th semester will be dedicated to writing the thesis and applying for jobs.

Activities	S1	S2	S3	S4	S5	S6
Bibliography	x	x	.	.	.	.
Axis 1:	x	x	x	.	.	
Axis 2:			x	x	x	
Axis 3:		x	x	x	x	.
Writing						x

Supervision and Research Environment

The Ph.D. thesis will be supervised by Pr. Sylvain Lamprier (LERIA-Universy of Angers) co-supervised by Ludovic Denoyer (Ubisoft, HDR). The Ph.D. student will share his time between Ubisoft (60%) and LERIA (40%). Since Ubisoft and LERIA are not located in the same city, we expect the student to come to LERIA one week each month, and will adapt this planing depending on the student constraints. This will be considered in the thesis budget to handle the co-location of the student.

This thesis work will be carried out in close collaboration between the company Ubisoft (industrial partner), in the La Forge team which is the R&D structure at Ubisoft, and the ARC Team of LERIA, with specializations in machine learning.

- In the ARC Team, Sylvain Lamprier (Full professor - LERIA), is specialized in deep learning and reinforcement learning. He has published in many high-venue international conferences (Neurips, ICML, ICLR, etc.) and selected journals as well (JMLR, Machine Learning, etc.). He has co-supervised 12 PhD students and is currently co-supervising 4 PhD.
- At Ubisoft La Forge, Ludovic Denoyer (HDR, full professor on sabatical, H-index 32) is co-author of more than 150 publications and has supervised about 20 PhDs. He has then worked as a research scientist at Facebook AI Research (FAIR) and joined Ubisoft in March 2022 to develop research axis toward the development of better video games.

4 Related Work

The topic of the thesis is related to *Imitation Learning* and *Inverse Reinforcement Learning* since it aims at learning from traces collected when real players are playing the game. In these domains, many methods have been proposed, starting from behavioral cloning (BC) [BS95], to more robust approaches like inverse reinforcement based on the maximum entropy principle [ZMBD08, FLA16], generative adversarial models [HE16], density ratio estimation [KNT19] or, more recently, Decision Transformers [CLR+21]. But these methods usually assume that the dataset is composed of expert traces, while in our case, the dataset is composed of multiple players' - suboptimal - traces. Works such as [BGNN19] attempt to achieve better-than-demonstration performances, assuming that experts are suboptimal. However, our setting is different, since we do not expect to maximize a given reward function, but to capture and control human play styles.

More related to our proposal, and particularly for axis 2 are the *one shot imitation learning* methods [DAS+17, DG21, DPC21] that aim at being able to infer a policy from a single trajectory. This has been declined in different contexts, also under the *meta-reinforcement learning* topic [YFX+18, YYFE19]. In our proposal, we face a slightly different problem where the objective is not to learn from one episode, but to learn from an incomplete episode, with variable length, such that the bot will replace the human during the game. Our objective share common properties with recent works in the

generative domain applied to sequential data [XXN+20, STZ21, CSK+22] from which we will get inspiration. One innovative direction that we envision is to ensure matching of style discrimination from available observations and fully generated trajectories using multi-dimensional reward-to-go in BC-like approaches. Another breakthrough will concern the matching of game-specific metric distributions between real-world observations and generated simulations.

At last, introducing the notion of controllability is very recent in the reinforcement learning literature, but much more frequent in order domains like natural language [XCC20, PBS20, ZSL+22], image [LZU+17, **?**] or video [FDC+20, DFLG21] generation. We expect to take inspiration from these methods to propose new solutions in reinforcement learning. In the RL fields, some recent methods, particularly based on diffusion models have emerged very recently [**?**, **?**]. But these methods do not include the style dimension and the difficulty will be to find ways to control a policy both using the play style and the different control dimensions.

References

[AD21] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

[BGNN19] Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pages 783–792. PMLR, 2019.

[BS95] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129, 1995.

[CLR+21] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[CSK+22] Jen-Hao Rick Chang, Ashish Shrivastava, Hema Koppula, Xiaoshuai Zhang, and Oncel Tuzel. Style equalization: Unsupervised learning of controllable generative sequence models. In *International Conference on Machine Learning*, pages 2917–2937. PMLR, 2022.

[DAS+17] Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. *Advances in neural information processing systems*, 30, 2017.

[DFLG21] Jérémie Donà, Jean-Yves Franceschi, Sylvain Lamprier, and Patrick Gallinari. Pde-driven spatiotemporal disentanglement. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[DG21] Sudeep Dasari and Abhinav Gupta. Transformers for one-shot visual imitation. In *Conference on Robot Learning*, pages 2071–2084. PMLR, 2021.

[DPC21] Christopher R Dance, Julien Perez, and Théo Cachet. Conditioned reinforcement learning for few-shot imitation. In *International Conference on Machine Learning*, pages 2376–2387. PMLR, 2021.

[FDC+20] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. *CoRR*, abs/2002.09219, 2020.

[FLA16] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. *CoRR*, abs/1603.00448, 2016.

[HE16] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *CoRR*, abs/1606.03476, 2016.

[KNT19] Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

[LZU+17] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5967–5976, 2017.

[PBS20] Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. Exploring controllable text generation techniques. *arXiv preprint arXiv:2005.01822*, 2020.

[STZ21] Lingfeng Sun, Masayoshi Tomizuka, and Wei Zhan. Multi-style human motion prediction and generation via meta-learning. 2021.

[XCC20] Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. On variational learning of controllable representations for text without supervision. In *International Conference on Machine Learning*, pages 10534–10543. PMLR, 2020.

[XXN+20] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, Xiaolong Wang, and Trevor Darrell. Hierarchical style-based networks for motion synthesis. In *European conference on computer vision*, pages 178–194. Springer, 2020.

[YFX+18] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*, 2018.

[YYFE19] Lantao Yu, Tianhe Yu, Chelsea Finn, and Stefano Ermon. Meta-inverse reinforcement learning with probabilistic context variables. *Advances in neural information processing systems*, 32, 2019.

[ZMBD08] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. 2008.

[ZSL+22] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*, 2022.