5 J. Liu, J. Togelius, D. Pérez-Liébana, and S. Lucas. Evolving Game Skill-Depth using General Video Game AI Agents. In *IEEE Congress on Evolutionary Computation* (CEC'17), 2017.

4.19 Explainable AI for Designers

Jichen Zhu (Drexel Univ. - Philadelphia, US), Rafael Bidarra (TU Delft, NL), Alex J. Champandard (AiGameDev.com KG - Wien, AT), Simon Colton (Falmouth University, GB), Reynald Francois (Ubisoft - Düsseldorf, DE), Matthew J. Guzdial (Georgia Institute of Technology - Atlanta, US), Amy K. Hoover (NJIT - Newark, US), Antonios Liapis (University of Malta - Msida, MT), Sebastian Risi (IT University of Copenhagen, DK), Gillian Smith (Worcester Polytechnic Institute, US), Anne Sullivan (University of Central Florida - Orlando, US), and G. Michael Youngblood (Xerox PARC - Palo Alto, US)

In response to the rapid technological success in AI, the emerging research area of Explainable AI aims to better communicate AI systems' decisions and actions to human users. The central goals of explainable AI are often to increase users' understanding, foster trust, and improve their ability to utilize the systems.

Explainable AI for designers (XAID), in particular, focuses on enhancing designers' capability to (co-)create user experiences with AI. Through the vantage point of computer games, we examine 1) the design space of explainable AI for game designers, 2) three case studies of XAIDs, and 3) design guidelines and open challenges in each case.

We identified different types of XAID techniques that can facilitate the game design and development process. In a broad stroke, they can be categorized into i) what to explain, ii) when to explain, and iii) how to explain. In terms of what to explain, XAIDs can be used to communicate many aspects of game AI. For instance, they can be used to explain the process of the chain of actions and reactions taken by game AI. Alternatively, they can simply explain the results of processing certain inputs. Regarding when to explain, the description may take place before, during, or after AI's operations, each scenario affording different types of explanations. As for how to explain, factors such as the form of explanations (e.g., as a tutorial or as justifications of specific AI actions at hand), levels of abstraction (e.g., concrete details or high-level abstraction), and the interaction model (e.g., AI as a tool for the human designers or as a co-creator with the designers) directly influences of how XAIDs should be designed.

To ground our survey of the XAID space, we conducted three case studies based on the type of AI systems (black-box or white-box) and the part of the game development process in which AI techniques are used. The case studies include XAIDs for 1) white-box procedural content generation (PCG) systems, 2) black-box PCG systems, and 3) black-box NPC behavior control.

In a white-box PCG system, we assume a system that has full knowledge of the underlying processes taking place; this allows XAID (e.g. in the form of text generation) to be embedded within the content generation code. White box PCG can hypothetically output a narrative (as a sequence of sentences), with each sentence produced following each command or function call. Ideal generative architectures for this approach are pipelines, where a number of generative processes are "chained", each producing an intermediate result which is taken by

the next process as input and producing an enhanced result as its own output [1]. In such an architecture, generation of textual explanation need not be internalized within the codebase, instead assessing the intermediate results from each process along the pipeline. Explanations in a white box PCG system can be produced at any point in the generative process and in any degree of clarity (as the explanation subsystem can have full knowledge of the underlying logic or ways in which content quality is assessed). Due to these reasons, the challenge for XAID in white box PCG is how to handle the possibly vast volume of information that can be output by such a system. Presenting a compelling and intuitive narrative to the user regarding the choices taken by the system can be done:

- ***sequentially*** in the order that the system makes decisions. This explanation can follow some form of story structure which simulates e.g. the generative pipeline [2]. Work on story generation can be used to enhance how the connections are made between different time slices in the generation (e.g. via natural language processing [3]) so that the narrative is coherent and causal links are made obvious. This can be achieved by post-processing the generated sequence of sentences to introduce throwbacks to past generative decisions which affect future outcomes or to foreshadow how one decision early on affects the final outcome.
- as ***highlights*** of the generative process by filtering out and omitting less interesting points in the generated sentence structure. For this to happen, a number of evaluation mechanisms are needed to assess interestingness (will this be interesting to a human user?), clarity (will this be understandable by a human user?), or creativity (is this point a creative milestone [4] where the design shifts?) of the text or the underlying generative commands that prompted it. Therefore, it is necessary to have the whole narrative (sequence of sentences) before the most interesting points within it are chosen in a post-processing step.
- ***non-sequentially***, summarizing the explanation starting from the most important points regardless of whether those are performed first or last in the code. Indeed, it is possible to start by presenting a description (visual, textual, or otherwise) of the final artifact and backtracking some of its most interesting elements on points in the generative process where those happened. Moreover, tropes such as sports game summaries can be used as inspiration, presenting the main outcomes of the generative process first (as non-sequential highlights) followed by a longer form of the sequential narrative regarding how generation progressed from unformed to fully formed content.

An open challenge in providing useful XAID is how to fit the entire processes of the white-box system into something that is compact and yet sufficient for designers.

In a black-box PCG system, we specifically looked into what type of XAID would be useful for game designers at a AAA game studio. We determined an "AI as student" framework matched designers intuitions when it came to artificial intelligence and black box machine learning techniques. To fit this framework a potential agent would need to: (1) share a common language of design with the human designer, (2) communicate its current understanding of this language, and (3) update this understanding in response to designer feedback. We identified three areas of existing work that matched these requirements: zero and one shot learning, explainable AI, and active learning respectively. In practice this would look like a designer pre-defining content with certain tags and an AI training on these tags (zero and one shot learning), the designer interrogating the agent's output content and tags through checking the maximally activated filters (explainable AI), and giving feedback through single examples, which the AI could use to retrain (active learning). A key **open challenge** in this area is how to filter out common sense knowledge in the explanation.

Finally, in a *black-box NPC behavior* control system, we focused on agents controlled by deep neural networks (DNNs), especially deep Q-networks (DQNs). Two types of information stood out as of particular importance to be well communicated to game designers. First, given a particular gameplay context, what is the likely distribution of actions the NPC can take at this given moment. Second, given a particular NPC action, what are all the possible situations that can lead to this action. Given the large number of possible actions and/or situations, similar to highlights in white-box PCG systems, a good design guideline for XAID is to highlight the unexpected and reduce the visibility of the common. A key open challenge to provide both types of information to a human designer is how to design the reward function.

In conclusion, explainable AI for designers is crucial for advancements in AI to be fully utilized in computer games and other types of interactive experiences. Results from this XAID workshop show that the current understanding of how to communicate the underlying AI algorithms to human designers is still rudimentary. Although different types of AI algorithms place varying challenges and opportunities for the corresponding XAIDs, an emergent key challenge shared in all three cases we investigated is **salience**. That is, among the various types of information that could be provided about the underlying AI algorithms, how do we define, identify, and communicate what is noteworthy. Future work includes deeper understandings of salience grounded in the specific needs of designers, algorithmic investigations of how to procedurally identify salient features, as well as design innovation of how to communicate them to human designers.

References

1 John Charnley, Simon Colton, Maria Teresa Llano Rodriguez and Joseph Corneli. The FloWr Online Platform Automated Programming and Computational Creativity as a Service. In *Proceedings of the International Conference on Computational Creativity*, 2016

2 Simon Colton, Jakob Halskov, Dan Ventura, Ian Gouldstone, Michael Cook and Blanca Perez-Ferrer. The Painting Fool Sees! New Projects with the Automated Painter. In *Proceedings of the International Conference on Computational Creativity*, 2015.

3 François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. In *Artificial Intelligence* 173, 7–8 (May 2009), 789–816. 2009.

4 Georgios N. Yannakakis, Antonios Liapis, Constantine Alexopoulos. Mixed-Initiative Co-Creativity. In *Proceedings of the 9th Conference on the Foundations of Digital Games*, 2014.

5 Panel discussions

5.1 Evaluation

Pieter Spronck (Tilburg University, NL)

At the end of the Seminar, we had a one-hour session with all attendees still present to evaluate the seminar. While the general consensus was that the Seminar had been a lot of fun and a great success, a desire was expressed to be a slightly less free-form for a future Seminar. In particular, the following points were brought up:

- Several attendees felt that they should have been primed more before arriving on the purpose of the Seminar and what was expected of them.