Guest Editorial: Special Issue on Human Centered AI in Game Evaluation

I. Introduction

ONE of the main objectives with which games are developed is to evoke positive experiences in their players. The purpose behind this objective ranges from pure entertainment or competition to fostering self-expression, social critique, or even knowledge exploration. Regardless of their specific aims, all games share an important phase—playtesting—a process aimed at ensuring that the player experience intended by design is achieved in the game.

On the one hand, playtesting is often both costly and time-consuming, as it typically relies on human players who may not be able to thoroughly evaluate every aspect of the game before release. Human testers can also experience fatigue and bring personal biases into the process. On the other hand, recent advancements in artificial intelligence (AI) and autonomous gameplay agents suggest the potential to use AI-driven players to evaluate game content quickly and affordably.

Researchers have worked on several methods aiming to support game designers, including the use of AI agents for automatic game testing and content evaluation. These techniques often create an extensive body of playtest data, which offer valuable insights to designers. Nevertheless, there is still a need to refine these methods to effectively extract the most meaningful insights. The review of existing research literature shows a lack of generalizability and verification in these approaches, as many studies focus on a single game or genre, without incorporating player modeling or experimental validation. For instance, AI-generated play traces may not accurately replicate human player behavior, which can lead to misinterpretations in results and artifact evaluations.

In this Special Issue, our objective was to focus on human-centered AI approaches, aiming for a more holistic and systematic approach to game evaluation. Putting the human player into the focus of the evaluation method may be beneficial for multiple reasons. First, evaluation systems must align with the designer's goals instead of assumed target measures. At the same time, other human stakeholders should also be considered, such as community managers, streaming communities, and esports organizations. Second, there is a need to bridge the gap between traditional human-centered evaluation (with human playtesters) and AI-driven methods: playtests are expensive, and AI methods often fail to account for human aspects, such as diverse player experiences, goals, or preferences.

A third reason is an aspect of generalizability. Current methods are typically only demonstrated in a single context; to improve generalizability, we need methods that work well across varied evaluation use cases. This includes the expansion into game evaluation for different genres, including mediums, such as tabletop games, and those integrating sensors, AR, or VR. Fourth, there is the need for designing AI tools and interfaces to be accessible, so that they can be easily and efficiently used by those who need them. This includes the need to investigate new evaluation metrics for game interfaces and control, extending the use of playtesting and game evaluation from the domain of games human–computer interaction to their application in related fields, such as procedural content generation (PCG) and game balancing.

The following section introduces briefly the 11 papers published in this issue. They cover a wide range of topics from player modeling (e.g., for meta or outcome prediction and playtesting) to the analysis and modeling of human player responses to aid designers (features include engagement, cognitive load, fear, and cultural differences). The editors would like to thank all authors for their submissions and all reviewers for their voluntary work providing feedback and insightful comments.

II. Publications in This Special Issue

Pan et al. [A1], propose a multimodal deep learning model for estimating the engagement levels of game streamers. They use facial, pixel, and audio information from videos taken from multiple players as input for their model. The system is able to assign credit to each input type for the classification algorithm, leading to the conclusion that the audio input is a key modality for engagement estimation. Their findings have potential applications in the field of game streaming and analysis of their audience.

Pan et al. [A2], conduct an AI-driven analysis of player reviews of souls-like games focused on assessing cross-cultural differences. The authors analyze reviews of 17 souls-like games in English, Chinese, and Russian. They collect these reviews from the Steam store and analyze them with natural language processing techniques, such as topic modeling and sentiment analysis. The findings include significant differences between player groups regarding play behaviors, opinions, and review sentiments. The paper is a nice example of human-centered evaluation that applies AI to better understand human perspectives on games.

Date of current version 17 December 2024.
Digital Object Identifier 10.1109/TG.2024.3507232

Sun et al. [A3], use different machine learning models to analyze data from wargame matches and predict game outcomes. The data are analyzed at a micro and macro level, aiming to capture player strategy and tactics. Their experimental work shows a high prediction accuracy, and it is complemented with an analysis on the different strategies followed by the players, both at a microlevel and a macrolevel.

Aylagas et al. [A4], propose a controllable PCG via machine learning method for generating Match-3 levels. They employ a variational autoencoder, which is conditioned on gameplay statistics and visual features. This way, they can generate levels with specific characteristics, such as symmetry or a desired average for the number of moves taken to finish the level. They evaluate the generated levels based on their adherence to the desired conditions, as well as using a domain-specific validity measure (levels need to be solvable within 20 moves).

Azizi and Zaman [A5], address the problem of automatic bug detection in games using a long short-term memory network to detect anomalies, and subsequently cluster bugs to determine their nature. In addition, the authors use deep Q-learning to simulate play-testing agents to gather datasets for bug detection, minimizing the requirement of human players. These models are tested on two first person shooters and one role-playing game.

Nam et al. [A6], generate levels for the game *Super Mario Bros.* employing the PCG via reinforcement learning (RL) framework. The generated levels are of diverse nature and at an adequate level of challenge for human players. The latter is achieved by evaluating these levels with a *degraded* A* agent, for which failures on the execution of its actions have been introduced. Examples are delayed actions, hastening, repetitions in several frames, and the introduction of preferences for in-game actions, as observed in human players. The authors evaluate their approach with human participants, showing that the levels produced by their approach have the appropriate level of challenge, variety, and lack of repetition of patterns.

Saravanan and Guzdial [A7], introduce a framework for discovering prevalent player strategies of a given game (metagame). Their motivation is to provide a tool for game developers to prevent unintended consequences for the metagame after releasing balancing adjustments. They propose to use (a set of) agents with winrates close to an average human's to simulate games with the planned changes. The framework also includes a prediction model for the popularity of team composition, which may consider domain knowledge and historic data. The authors demonstrate and test the framework and its metagame predictions on historic data of Pokémon Showdown.

Ogawa et al. [A8], introduce a method for developing human-like gameplay model by blending supervised and RL approaches to improve move-matching accuracy in board games, such as *Chess, Go,* and *Shogi.* The proposed model, "Blend," combines the Maia model, based on supervised learning and trained on human game data, with an AlphaZero-inspired RL model, leveraging strengths from both to address the limitations of each approach. Notably, while supervised models often mimic human moves effectively, they may lack accuracy in less-represented scenarios; RL models, on the other hand, perform well in challenging positions but may deviate from human-like behavior. Experiments demonstrate that the Blend model enhances human-like gameplay, achieving higher move-matching accuracy across various player skill levels in different games. This approach not only improves the interactive quality of game AI but also offers insights for broader applications where AI requires human-like decision-making traits.

Gutiérrez-Sánchez et al. [A9], introduce a task-guided RL algorithm designed for interpretable automated regression testing in video games. They employ truncated linear temporal logic (TLTL) to create interpretable reward structures, which guide RL agents in performing design validations without requiring prior ML expertise. The approach generates dense reward functions directly from TLTL task specifications, facilitating a more straightforward and debuggable testing process. Through experiments in a 3-D game environment, the authors demonstrate that their method can adapt to design modifications in complex video game scenarios, highlighting its effectiveness for automated regression testing compared to traditional and imitation-based strategies.

Pretty et al. [A10], examine the effectiveness of various physiological indicators of cognitive load—specifically, electroencephalography, electromyography, heart rate, heart rate variability, electrodermal activity, and eye blink rate—in predicting players' subjective cognitive load during gameplay. They detail the factors to consider when selecting appropriate sensors that can reliably serve as proxies for the cognitive demand experienced by players.

Zhang et al. [A11], report on the experiences of players in VR horror games based on insights from 25 interviews. Specifically, they discuss the factors influencing fear, the strategies players use to cope with it, and the drivers of player behavior. Based on the findings, they provide some suggestions for more effective design of VR horror games.

ALENA DENISOVA, *Guest Editor*
University of York
YO10 5DD York, U.K.
alena.denisova@york.ac.uk

DIEGO PEREZ-LIEBANA, *Guest Editor*
Queen Mary University of London
E1 4NS London, U.K.
diego.perez@qmul.ac.uk

VANESSA VOLZ, *Guest Editor*
Centrum Wiskunde & Informatica (CWI)
1090 GB Amsterdam, The Netherlands
vanessa.volz@cwi.nl

JULIAN FROMMEL, *Guest Editor*
Utrecht University
3584 CS Utrecht, The Netherlands
j.frommel@uu.nl

SAHAR ASADI, *Guest Editor*
King, Microsoft Gaming
111 65 Stockholm, Sweden
saharasadi@microsoft.com

Appendix: Related Articles

[A1] S. Pan, G. J. Xu , K. Guo, S. H. Park, and H. Ding, "Video-based engagement estimation of game streamers: An interpretable multimodal neural network approach," *IEEE Trans. Games*, vol. 16, no. 4, pp. 746–757, Dec. 2024,, doi: 10.1109/TG.2023.3348230.

[A2] S. Pan, G. J. Xu, K. Guo, S. H. Park, and H. Ding, "Cultural insights in souls-like games: Analyzing player behaviors, perspectives, and emotions across a multicultural context," *IEEE Trans. Games*, vol. 16, no. 4, pp. 758–769, Dec. 2024,, doi: 10.1109/TG.2024.3366239.

[A3] Y. Sun, Y. Sun, J. Yu, Y. Li, and X. Zhou, "Predicting wargame outcomes and evaluating player performance from an integrated strategic and operational perspective," *IEEE Trans. Games*, vol. 16, no. 4, pp. 770–782, Dec. 2024,, doi: 10.1109/TG.2024.3369330.

[A4] M. V. Aylagas, J. Bergdahl, J. Gillberg, A. Sestini, T. Tolstoy, and L. Gisslén, "Improving conditional level generation using automated validation in match-3 games," *IEEE Trans. Games*, vol. 16, no. 4, pp. 783–792, Dec. 2024,, doi: 10.1109/TG.2024.3440214.

[A5] E. Azizi and L. Zaman, "AstroBug: Automatic game bug detection using deep learning," *IEEE Trans. Games*, vol. 16, no. 4, pp. 793–806, Dec. 2024,, doi: 10.1109/TG.2024.3402626.

[A6] S. Nam, C.-H. Hsueh, P. Rerkjirattikal, and K. Ikeda, "Using reinforcement learning to generate levels of Super Mario Bros. with quality and diversity," *IEEE Trans. Games*, vol. 16, no. 4, pp. 807–820, Dec. 2024,, doi: 10.1109/TG.2024.3416472.

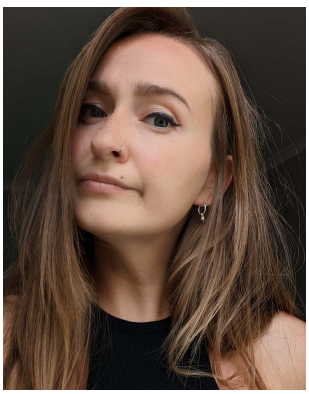
[A7] A. Saravanan and M. Guzdial, "A framework for predicting the impact of game balance changes through meta discovery," *IEEE Trans. Games*, vol. 16, no. 4, pp. 821–830, Dec. 2024,, doi: 10.1109/TG.2024.3457822.

[A8] T. Ogawa, C.-H. Hsueh, and K. Ikeda, "More human-like gameplay by blending policies from supervised and reinforcement learning," *IEEE Trans. Games*, vol. 16, no. 4, pp. 831–843, Dec. 2024,, doi: 10.1109/TG.2024.3424668.

[A9] P. Gutiérrez-Sánchez, M. A. Gómez-Martín, P. A. González-Calero, and P. P. Gómez-Martín, "A progress-based algorithm for interpretable reinforcement learning in regression testing," *IEEE Trans. Games*, vol. 16, no. 4, pp. 844–853, Dec. 2024,, doi: 10.1109/TG.2024.3426601.

[A10] E. J. Pretty, R. Guarese, C. A. Dziego, H. M. Fayek, and F. Zambetta, "Multimodal measurement of cognitive load in a video game context: A comparative study between subjective and objective metrics," *IEEE Trans. Games*, vol. 16, no. 4, pp. 854–867, Dec. 2024,, doi: 10.1109/TG.2024.3406723.

[A11] H. Zhang, X. Li, X. Fu, C. Qiu, J. Zhang, and J. M. Carroll, "Understanding fear responses and coping mechanisms in VR horror gaming: Insights from semi-structured interviews," *IEEE Trans. Games*, vol. 16, no. 4, pp. 868–881, Dec. 2024,, doi: 10.1109/TG.2024.3403768.



Alena Denisova received the Ph.D. degree in perception of adaptive technologies in video games from the University of York, York, U.K., in 2017. She is currently a Senior Lecturer (Associate Professor) with the Human-Computer Interaction Research Group, University of York, York, U.K. Her research interests include understanding and improving player experience of video games (e.g., perceived challenge) and developing and evaluating tools and methods for researching interactive experiences.

Diego Perez-Liebana received the Ph.D. degree, with focus on the application of Monte Carlo Tree Search and Rolling Horizon Evolution to real-time games, from the University of Essex, Colchester, U.K., in 2015. He is currently a Senior Lecturer with Queen Mary University of London, London, U.K. He is the Co-Founder of Tabletop R&D, a startup that provides AI-driven automatic testing services to tabletop board game designers. His research interests include the application of game-playing AI agents to tabletop board games, automatic playtesting, and AI-based game design.

Vanessa Volz received the Ph.D. degree, with focus on surrogate-assisted evolutionary algorithms with focus on their application to games, from TU Dortmund University, Dortmund, Germany, in 2019.

She is currently a Tenure Track Researcher with the Evolutionary Intelligence (EI) Group, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands. Her research focuses on transfer learning in the context of evolutionary computation, especially in the context of recurring or otherwise dynamic problems.

Julian Frommel received the Ph.D. degree in computer science, with a focus on player state assessment, from Ulm University, Ulm, Germany, in 2019. He is currently an Assistant Professor of interaction/multimedia with Utrecht University, Utrecht, The Netherlands. His research interests include the design and implementation of interactive digital systems that provide enjoyable, meaningful, safe, and healthy experiences for users, including research on how to mitigate the negative effects of toxicity and harassment in online games and other online spaces.

Sahar Asadi received the Ph.D. degree from Örebro University, Örebro, Sweden, in 2017. She is currently the Director of AI Labs, King (part of Microsoft), Stockholm, Sweden, where she leads AI research for mobile games. She has been conducting applied research on generalization and scalability of learning methods in various industry applications. Her research interests include reinforcement learning for playtesting, representation learning and player understanding, and responsible AI.