Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2025)

FighterDDA: A Simulation Testbed for Evaluating Director-Based Dynamic Balancing

Samuel Shields[1], **Edward F. Melcer**[2]
[1]University of California, Santa Cruz, Santa Cruz, CA
[2]Carleton University, Ottowa, ON
samshiel@ucsc.edu, edwardmelcer@cunet.carleton.ca

Abstract

While AI Directors have been widely studied, less attention has been paid to how different director styles can coexist in a single game environment with divergent design goals. This paper presents FighterDDA, a simulation testbed designed to evaluate dynamic balancing strategies aimed at different audiences using varied AI Directors in turn-based role-playing games. The testbed can rapidly simulate thousands of games, output formatted data, and provide visualizations of play traces for designers to evaluate whether a director strategy was successful. In an initial evaluation with the system, we found that a director made to even out win rates between differently-skilled players as well as a director to tune the overall difficulty for equally skilled players both showed success. The system demonstrates how a simulated testbed with data reporting supports prototyping and evaluating varied director systems.

Code — https://github.com/smshields/FighterDDA

Introduction

Video game balance is a multidimensional concept tied to perceived fairness (Schreiber and Romero 2021). Balance is key to sustaining long-term player engagement (Andrade et al. 2006), and has been frequently cited in the industry as key to keeping a game fresh and retaining a player base over a long period of time (Choudhury, Fujita, and Lo 2018; Krell 2021). Crucially, game balancing can be an expensive and time-consuming task, as it is generally required during most of a game's development lifecycle and requires extensive playtesting to confirm different hypotheses and approaches to achieve balance (Jeon et al. 2023; Kokkinakis et al. 2021).

The complexity of balance does not end at implementation — understanding desired player experience and audience composition plays a major role in what balancing strategies will work where (Fullerton, Swain, and Hoffman 2004). Even within a single game, there might be multiple balancing approaches employed to ensure that a game is accessible to new players while maintaining fair, skill-based competition for entrenched players. Understanding how to align AI systems with such varied audiences remains an open challenge.

In this work, we introduce our novel platform *FighterDDA* for exploring methods to overcome these challenges in game balance through the use of different AI directors (AID) and visualization tooling. Tools such as the one we present could help designers get closer to a balancing goal prior to testing, evaluate the success of a balancing approach from a quantitative perspective, and understand if a given design direction holds promise. We ground our system in the turn-based role-playing game (TBRPG) genre, using two different AIDs with varying goals. Namely, the directors balance for 1) audiences of similar skill levels and 2) audiences of differential skill levels. We found that our approach of working in a simulated environment containing multiple AID implementations alongside visualization tooling helped fulfill our targeted design goals.

AI Directors

AI Directors are a dynamic game balancing technique where an agent (often invisible to the player) watches and evaluates the current game state, providing changes to the runtime system to change the play experience to better align with designer goals (Foffano and Thue 2019). The *Hamlet* system introduced by Hunicke is an early documentation of this approach, changing item allocations dynamically in response to player progression (Hunicke 2005). A well-known version of this approach appears in the game *Left4Dead* (South 2008), where enemy spawns are driven by a perceived level of intensity in the gameplay (Booth 2009). Directors can vary depending on designer goals—e.g., leveling the field for mismatched players or modulating difficulty for evenly matched ones. Key elements of directors include the game state space and tuneable parameters (what does the director evaluate and use to change game state), the manager policy (how does the director change the game state), and the opportunities to act (when does a director act) (Thue and Bulitko 2018). These elements can be seen in context in figure 3. Using a director with multiple implementation methods is present in work surrounding the testbed *FarmQuest*, which uses PaSSAGE (Thue, Bulitko, and Spetch 2008) and reinforcement learning director approaches in regards to quest selection in a cozy game (Kristen, Guzdial, and Sturtevant 2022, 2024); though these works evaluate the efficacy of a single design strategy with multiple technical implementations instead of methods that target fundamentally different

design goals.

Turn-Based Role-Playing Games

TBRPGs have been a staple of video games ranging back to the 1980's, evidenced by franchises such as the *Ultima* series (Systems 1981). Descendants of tabletop role-playing games, TBRPGs have enjoyed commercial and critical success, ranging from the best-selling *Final Fantasy* series (Square 1987) to the 2025 critical darling *Clair Obscur: Expedition 33* (Interactive 2025). These games have common attributes: the player controls some number of characters with varying abilities (attack, heal, magic). Characters have some form of speed stat which determines turn order. Characters work together to defeat enemies, who possess their own sets of abilities. Characters win by reducing their opponents' health to zero. The popularity of these games, as well as their ability to be efficiently simulated headlessly makes them an ideal context in which to implement and test various director-based dynamic balancing systems. AIDs are underexplored in TBRPGs, presenting an opportunity to use a common yet under-examined environment to test director implementation. The genre's adversarial play, complex action space, and simplified action queue makes TBRPGs a good starting metaphor for other genres that layer on additional dimensions to this framework (e.g. dexterity in action/adventure RPGs).

System Description

Our system aims to evaluate director approaches in the context of a TBRPG. To generate sufficient data for balance analysis, we built a headless system that mimics real-time TBRPGs such as *Final Fantasy IV* (Square 1987). We implemented the system to be controllable by one or two players, enabling investigation into single- and multiplayer environments. All characters have unique archetypes (warrior, mage, cleric, and rogue) that align with established patterns in the RPG space. Each character's stats determine action damage, order, and effectiveness. Initial stat fuzziness introduces team variation between runs. The game ends when one team is eliminated or a maximum number of actions is reached (resulting in a draw). The full source code for the simulation and corresponding tooling can be found on GitHub (a link to repository can be found at the beginning of this work).

Simulation

The simulation of the game runs in time steps. Each step, character speeds are used to determine if a player can add an action to the queue. Additionally, the simulation checks if actions need to be executed from the queue or if the AID should perform an intervention. Simulations are seeded for reproducibility. Automated playtesting agents are controlled by utility agents as they are a pattern with demonstrated success in industry (Thompson 2025). These agents evaluate known game state (excluding hidden opponent data) and decide what action to perform for each character they control. Actions are scored to prioritize defeating the opponent while preserving team health. A weighted probability based on utility scores is used for final action selection. Within the current system, there are three agents that emulate different players:

1. **Optimal** — Selects higher-value actions. Emulates an experienced player trying to win games decisively.
2. **Random** — Selects actions randomly. Emulates a new player. Consistently loses to an optimal player.
3. **Griefer** — Selects lower-value actions. Emulates a player attempting to diminish play experience. Stalls games without attempting to win.

Directors

The directors are state-based goal-based agents that seek to accomplish a designer goal. In our simulation, we implement two such possible goals for our directors to operate on. In one condition (differential skill level), we seek to level the playing field for players of different skill levels in a multiplayer environment. The director accomplishes this by examining the stats of both teams independently, projecting potential damage trends as a result of stat changes, and then applying stat increases to the losing team and/or stat decreases to the winning team. This aims to close the win/loss gap between players. In the second director condition, we aim to modulate the overall difficulty of the game for one or two players by adding environment-oriented changes to scalars used in damage and healing calculations. This resembles environmental effects like "fog" that reduce damage output. This director monitors team health trends, fits a line to recent data, and compares it to a designer-defined ideal trend line. The distance between the next predicted points from the line of best fit and the ideal trend line determines the magnitude and direction of the director adjustment.

Data Logging and Visualization

Simulations can run and evaluate games in large batches up to approximately 10,000. Each game outputs a JSON file that includes starting parameters (e.g. targeted difficulty), logs each time step with actions, director interventions, character stat changes, and summarizing metrics. Separate logs provide a human-readable summary (also viewable in console; see figure 1). These logs are then parsed by a graphing utility (see figure 2) that enables turn-by-turn analysis of a game along with important summary metrics. The visualization, JSON game files, and overall simulation data output provide ample material for a designer to review how well a given AID implementation worked.

Evaluation

Our evaluation consisted of testing simulation output against two defined designer goals: 1) can a director be applied to lessen the gap between differently skilled players, and 2) can a director be applied to change the overall difficulty for two similarly skilled players as reflected in game length. To answer these questions, we simulated 1000 games per director configuration. In the differential skill condition, the director significantly reduced the win/loss/draw disparity, either increasing the lower-skilled player's win rate or increasing

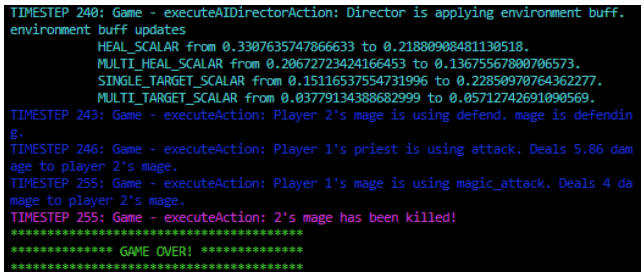

Figure 1: Console output detailing game actions and results.

Condition	Director Disabled			Director Enabled		
	P1	P2	Tie	P1	P2	Tie
Optimal, Optimal	47.5%	49.6%	3%	46.4%	45.5%	8.1%
Random, Random	42.7%	38.1%	19.2%	17.4%	18.2%	64%
Griefer, Griefer	0%	0%	100%	0%	0%	100%
Optimal, Random	93.3%	5.3%	1%	59.1%	16.8%	24.1%
Optimal, Griefer	55.2%	0%	44.8%	36.7%	0%	62.3%
Random, Griefer	30.8%	0%	69.2%	9.5%	0.2%	90.3%

Table 1: Player Win Rates Using Differential Skill-Level Director

the draw rate between players. Table **??** details the results. In the similar skill condition, we found that setting a target difficulty could significantly impact game length while maintaining consistent win rates and avoiding draws. The average number of actions per game based on difficulty setting ranged from 28.55 to 76.96. These results suggest both directors align with their intended design goals.

Discussion

The importance of this work is twofold—on the one hand, AIDs can achieve diverse balancing goals through targeted implementation. On the other hand, thorough data reporting, visualization, and a high number of simulations can be used to increase confidence that a given approach works as intended. These insights suggest how such systems could streamline the costly, time-intensive balancing process. Our cursory evaluation showed that a director framework combined with detailed analytics can support varied design goals. The investigation of how directors can be designed for fundamentally different use cases and player audiences helps clarify how analytics and visualization can inform AID strategies.

Generating Macro- and Micro-Gameplay Data for Balancing

The system's produced data are useful for two main reasons - it enables a macro-level view of a large number of simulations while enabling a micro-level view of individual play traces. In addition to the critical stats mentioned in the evaluation, tooling allows designers to inspect play traces from any seed. Our initial metrics are relatively simple and don't provide a realistic judgment of if the balancing achieved actually matches desired gameplay style (is the encounter a weak enemy, a boss, etc.). However, getting to roughly the right target provides a generated set of play traces that can be examined for specific desired patterns of behavior. For example, a designer might be seeking a sense of drama in games, looking for situations where players are constantly exchanging the lead between one another. These can be identified both by metrics and the visual tooling provided in figure 2. Ensuring macro-level consistency while meeting micro-level goals provides the designer a starting point to confirm their hypotheses with human playtesters.

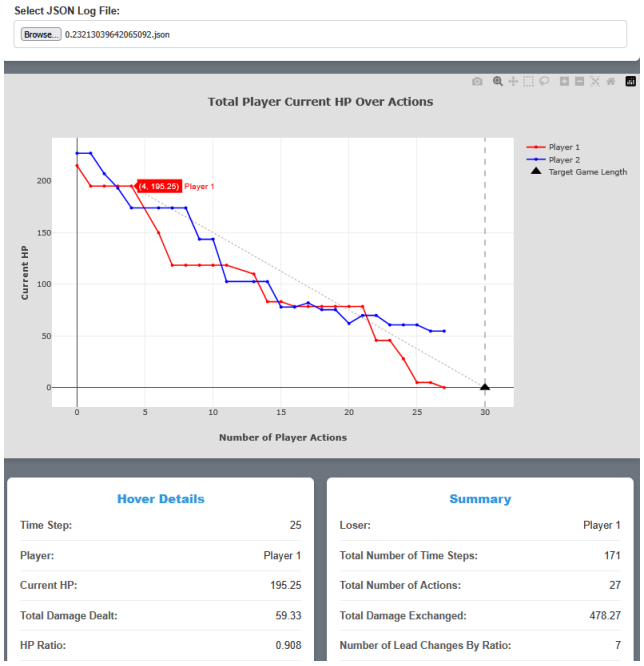

Figure 2: Tooling included with the simulation system to inspect individual game play traces from a batch of simulations. The graph details both players' current health, while allowing inspection of game state at each action performed. A trend line shows the slope desired for a game's difficulty (if specified).

Future Work and Limitations

This work-in-progress system's output shows promise for future design and analytics tools. A next step for the system would be to implement some form of expressive range analysis, which would enable a designer to see how successful a given metric is while also visualizing the range of games possible under given parameters (Withington and Tokarchuk 2023). A playable version of the system would enable two types of user studies - player evaluation to understand if the different balancing approaches are successful for their intended audiences, and a designer survey to assess the tool's usefulness. Several limitations remain for future work—the player agents specified do not learn over time, meaning they

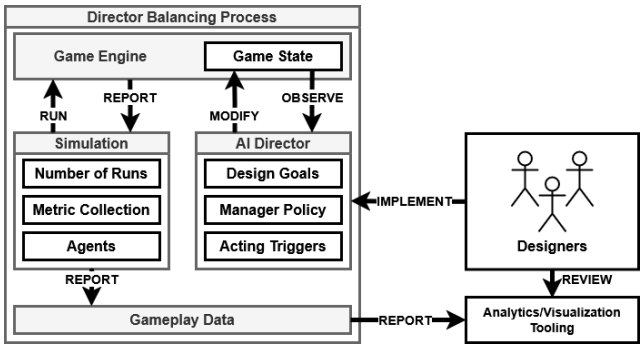

Figure 3: A conceptual diagram detailing the interaction between a game engine, simulation manager, director manager, data collection, and designer approaches. Our tooling suggests an iterative loop using the visualized output data from a simulation involving a director to make tuning adjustments to the director itself.

have a disconnect from players who would change their strategy during gameplay. Implementing a learning agent could address this issue. In addition, the initial evaluation used relatively simple markers for the success of a director. In future work, better identifying metrics linked to player satisfaction would help refine AID strategy and evaluation.

Conclusion

Combining reporting with player-focused AIDs offers an iterative way for designers to evaluate how their AID systems might achieve their balancing goals. We found through an initial evaluation that both diverse director approaches and thorough data reporting could expedite balancing, enabling more refined prototypes before user testing.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2202521. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Special thanks to the Expressive Intelligence Studio at UC Santa Cruz and Alternative Learning Technology lab at Carleton for allowing me to demo this work while providing useful feedback. Thanks also to Oliver Withington for help debugging and suggesting improvements to the system.

References

Andrade, G.; Ramalho, G.; Gomes, A.; and Corruble, V. 2006. Dynamic game balancing: An evaluation of user satisfaction. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 2, 3–8.

Booth, M. 2009. The ai systems of left 4 dead. In *Artificial Intelligence and Interactive Digital Entertainment Conference at Stanford, 2009*.

Choudhury, S.; Fujita, A.; and Lo, B. 2018. CEO of video game maker Nexon flags the 'single most important idea' in his industry — cnbc.com. https://www.cnbc.com/2018/03/19/video-games-nexon-ceo-says-longevity-is-critical-for-video-games.html. [Accessed 13-09-2024].

Foffano, F.; and Thue, D. 2019. Changes of user experience in an adaptive game: a study of an AI manager. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, 1–8.

Fullerton, T.; Swain, C.; and Hoffman, S. 2004. *Game design workshop: Designing, prototyping, & playtesting games*. CRC Press.

Hunicke, R. 2005. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, 429–433.

Interactive, S. 2025. Clair Obscur: Expedition 33. [PlayStation 5, Windows, Xbox Series X/S].

Jeon, H.-C.; Baek, I.-C.; Bae, C.-m.; Park, T.; You, W.; Ha, T.; Jung, H.; Noh, J.; Oh, S.; and Kim, K.-J. 2023. RaidEnv: Exploring new challenges in automated content balancing for boss raid games. *IEEE Transactions on Games*.

Kokkinakis, A.; York, P.; Patra, M. S.; Robertson, J.; Kirman, B.; Coates, A.; Chitayat, A. P. P.; Demediuk, S.; Drachen, A.; Hook, J.; et al. 2021. Metagaming and metagames in Esports. *International Journal of Esports*, 1(1).

Krell, J. 2021. How does Riot Games balance 'League of Legends?' The answer is a bit meta. https://www.washingtonpost.com/video-games/esports/2021/02/22/league-legends-meta-riot/. [Accessed 13-09-2024].

Kristen, K. Y.; Guzdial, M.; and Sturtevant, N. R. 2022. Farmquest: A demonstration of an ai director video game test bed. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, 288–290.

Kristen, K. Y.; Guzdial, M.; and Sturtevant, N. R. 2024. The FarmQuest Player Telemetry Dataset: Playthrough Data of a Cozy Farming Game. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, 263–268.

Schreiber, I.; and Romero, B. 2021. *Game balance*. CRC Press.

South, V. 2008. Left 4 Dead. [Windows, Xbox 360, macOS].

Square. 1987. Final Fantasy. [Nintendo Entertainment System].

Systems, O. 1981. Ultima. [Amiga].

Thompson, T. 2025. AI 101: Introducing Utility AI — aiandgames.com. https://www.aiandgames.com/p/ai-101-introducing-utility-ai. [Accessed 31-05-2025].

Thue, D.; and Bulitko, V. 2018. Toward a unified understanding of experience management. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14, 130–136.

Thue, D.; Bulitko, V.; and Spetch, M. S. 2008. PaSSAGE: A demonstration of player modelling in interactive storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 4, 226–227.

Withington, O.; and Tokarchuk, L. 2023. The right variety: Improving expressive range analysis with metric selection methods. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, 1–11.