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

Modeling the Internal Monologue: A Framework for Reconciling Authorial Intent with Agent Autonomy

Kyle Mitchell

University of California, Davis
1 Shields Avenue, Davis, California 95616
kdmitch@ucdavis.edu

Abstract

The design of narrative agents is often caught in a trade-off between authorial control and character agency. This work proposes the existence of a middle-ground approach, arguing that a robust, authored narrative can emerge from the performative autonomy of deeply simulated, coherent agents. To explore this thesis, I have developed a novel, three-layer architecture that synthesizes Stanislavskian performance theory, defeasible logic for argumentation-based reasoning, and a dynamic behavior language for execution. My work to date includes implementing this architecture for goal attainment in the *Oops! All Bards* prototype, and exploring foundational personality models for goal formation in the *Stay Thy Blade* prototype. The remaining research directions presented for feedback include the synthesis of these two tracks, the development of a constrained knowledge authoring framework, and a multi-pronged evaluation strategy. The contribution of this dissertation will be a validated framework that attempts to bridge the gap between agent autonomy and authorial intent, enabling the creation of "strong stories through strong characters."

Introduction

My personal research agenda is driven by a desire to create virtual agents whose actions are the performative result of a simulated internal monologue. The modeling of this transparent, character-centric deliberation is often deprioritized in agent architectures, as dominant paradigms embed logic in procedural structures like Behavior Trees (Colledanchise and Ögren 2018), focus on plan validity over nuanced justification like Goal-Oriented Action Planning (Orkin 2006), or rely on deep but opaque models seen in utility-based systems, machine learning, and social simulation (Świechowski, Szymański, and Mańdziuk 2023; Alifieris, Siolas, and Stafylopatis 2023; McCoy et al. 2011). This leaves a gap for architectures that use formal argumentation as the core mechanism for an agent's internal monologue.

My approach is grounded in Stanislavski's performance theory, where a character's high-level goal is a supertask pursued through context-dependent behaviors (Stanislavski 1989; Stanislavski and Hapgood 2012; Stanislavski 2013). The selection of these behaviors is inspired by Stanislavski's "Magic If," a process of justification through inquiry where an agent queries its knowledge base to find a character-driven reason to act. I implement this framework through *Viv*, a novel three-layer architecture. Its key innovation is a Defeasible Logic Programming (DeLP) reasoning layer (García and Simari 2004; Antoniou et al. 2000; Agis, Cohen, and Martínez 2016) that provides a formal mechanism for this internal monologue, informing a flexible tactical layer built with A Behavior Language (ABL) (Mateas and Stern 2004, 2002). This architecture is currently being tested in a prototype game, *Oops! All Bards* (OAB).

A second, complementary track of my research explores the distinct but related problem of dynamic goal formation. For this, I am developing components in isolation within a 2D prototype, *Stay Thy Blade* (STB), that use character personality models to guide the autonomous formation of a character's high-level goals, or "supertasks." My dissertation will synthesize these two efforts, presenting a framework for agents that can both form their own goals and pursue them through an explainable, performative process. This will involve tackling future challenges, such as researching how personality traits can influence the dialectical reasoning process itself and exploring how to computationally represent a character's entire journey, not just a single objective. This paper details the progress on both fronts and outlines the remaining work.

Architectural and Applied Contributions

As stated, my research contributions to date have unfolded along a few complementary thrusts: the design of a complete architecture for goal attainment, a testbed game environment in which this system can be reasonably assessed, and a foundational exploration of personality modeling for future work in goal formation.

The Viv Architecture for Goal Attainment

My primary work has been on the problem of goal attainment, addressed by the design and implementation of *Viv*, a novel three-layer agent architecture (Mitchell and McCoy 2024). The strategic layer, implemented in C# within the Unity engine, operationalizes Stanislavskian acting principles. A singleton *Viv* manager orchestrates the process, registering all *VivCharacter* components in the scene. Each

VivCharacter acts as a data hub for an agent, holding references to *ScriptableObject* assets that define its knowledge (*DELPEntity*) and personality (*VivPersona*), a design choice made to facilitate a designer-friendly workflow. This strategic layer is served by the reasoning layer, which runs as an external Java process using Defeasible Logic Programming (DeLP) to provide a traceable "internal monologue" of competing arguments. We chose defeasible logic over standard monotonic logic because it provides a formal mechanism for reasoning with the incomplete and conflicting information characteristic of believable characters. Furthermore, this symbolic, "glass box" approach was deliberately chosen to prioritize a transparent reasoning process, distinguishing it from modern generative approaches whose justifications for action are often implicit and difficult to trace. The resulting plan is then dispatched as a *VivWME* to the tactical layer, which uses A Behavior Language (ABL) (Mateas and Stern 2004) on the same Java server. This layer employs a "hivemind-daemon" pattern, allowing a shared library of behaviors to be reused by all characters.

Testbed Game Environment

To prove the viability of this architecture, I have implemented it within my primary prototype, a classic, D&D-style roleplaying game (RPG) titled *Oops! All Bards* (OAB) (Mitchell, Pettijohn, and McCoy 2022). In OAB, the player creates a character from a selection of performance-themed classes and is set within a medieval fantasy world. Its RPG structure serves as an effective testbed, as it provides a clear, authored narrative spine through its quests, but also offers the player significant freedom to explore, interact, and potentially deviate from the expected path. This creates the kinds of scenarios where a purely scripted character would fail. The *Viv* architecture can be used to control a wide range of non-player characters (NPCs) in OAB, from antagonists to companions who can join the player's "Band." This design allows me to test the system's core capabilities in a live, interactive environment. For instance, the *Viv*-controlled "guild enforcer" NPC, Wurguth, must react to the player's unscripted investigations in the sleepy starting village of the game. His behavior dynamically shifts between measured confrontation and outright hostility, based not on a pre-authored branching path, but on the new facts added to his knowledge base by the game's *KnowledgeUpdateRule* system.

Personality Modeling for Goal Formation

My secondary line of work explores a foundational component required for future goal formation research. Here, my goal was to establish a basis for creating and representing character personalities that could eventually drive goal selection. My approach was to develop and test this concept in isolation within a simplified, interactive version of Shakespeare's *Macbeth*, a prototype titled *Stay Thy Blade* (STB) (Treynor et al. 2025), in which the player is asked to step in during a pivotal moment of the play to attempt to soothe the "scorpions" of Macbeth's mind, and talk him down from committing violent acts. The social and political landscape of *Macbeth* provides a rich testbed for modeling complex motivations. In this particular prototype, the character of Macbeth is authored with a set of personality traits (e.g., ambition, guilt) which directly influence their set of low-level boolean beliefs relevant to conversation. This prototype serves as a successful proof-of-concept, demonstrating a necessary first step toward my long-term goal of synthesizing this work with the *Viv* architecture. The vision is for an agent whose foundational personality traits, as explored in STB, can dynamically drive the formation of its high-level supertasks, which it would then pursue using *Viv*'s explainable goal-attainment process.

Proposed Research Directions

The work completed to date provides a strong foundation for the remaining contributions of my dissertation, which are organized into the following research thrusts.

Synthesizing Personality and Reasoning

How can innate character personality traits be modeled to directly influence the process of formal, argumentative reasoning, moving beyond simply sorting behavioral outcomes? While my prior work in a prototype called *Sunset Valley* showed that traits can influence social behaviors like gossiping (Liao et al. 2023), and influential systems like *The Sims 3* and *Versu* (Evans and Short 2014) use personality to select from pre-authored actions, this research focuses on how personality can alter the formal justification process itself. Building upon foundational personality research (McCrae and Costa 1992; Shirvani 2021), my proposed work will investigate several potential mechanisms for this synthesis. These include having traits alter the strength of defeasible rules, dynamically add or retract entire sets of rules (e.g., a "paranoid" trait activating suspicion rules), or influence how an agent interprets uncertainty (e.g., an "optimistic" character treating an "UNDECIDED" query as a "YES").

The Abstract Supertask

The synthesis described above is a step toward addressing a more foundational challenge, which forms my second research question: What computational structures, beyond a static goal, are required to model a character's long-term journey, or "supertask," as described in Stanislavskian theory? To explore this, my research proposes extending the *VivPersona* from a simple priority list into a rich, dynamic character model. This extended persona would incorporate data structures for tracking an agent's core values, its long-term memories of key narrative events, and its evolving social relationships with other characters. The objective of this research thrust is to create a system where a character's supertask is not assigned, but emerges organically from this rich, persistent model of their inner life.

Knowledge Engineering and Domain Authoring

A challenge in the current architecture is that the expressive power of DeLP is effectively limitless. This leads to a related research question: How can this expressive power be constrained into a practical, reusable authoring framework for

game designers? Without a constrained framework, authoring a consistent and maintainable knowledge base requires a high degree of expertise in logic. My proposed work will address this by researching and developing a domain-specific ontology for a classic RPG. This would act as a "design bible" or template (Schell 2019), codifying the core predicates and rule structures for common concepts like social relationships, character reputation, and location properties. The goal is to create a reusable starting point that abstracts away much of the underlying logical complexity, lowering the barrier to entry for users of the system.

Evaluation Strategies

Finally, the completed dissertation will present a multi-pronged evaluation of the synthesized architecture. This evaluation is designed to answer three specific research questions:

1. How does this architecture's authoring workflow and runtime performance compare to established paradigms?
2. What are the performance bottlenecks of the client-server reasoning architecture and how does it scale?
3. Do players perceive characters driven by this architecture as more coherent, explainable, or believable?

To address these questions, I will employ the following methods:

Comparative Implementation Study A comparative analysis will be conducted against Behavior Trees and GOAP. This will involve re-implementing a complex scenario from *Oops! All Bards* in each architecture and measuring quantitative metrics (e.g., lines of code and number of files modified to author a new behavior) as well as qualitative authoring heuristics. Runtime performance metrics such as CPU usage and decision latency per agent will also be compared.

Technical Scalability Testing A series of stress tests will be conducted to evaluate the system's performance at scale. This involves programmatically increasing the number of concurrent agents controlled by the proposed architecture within the OAB prototype while measuring performance indicators like server response time, total memory load, and any degradation in individual agent deliberation speed.

Qualitative User Study A user study with a between-subjects design will be conducted to address the question of player perception. One group of participants will play a scenario in the OAB prototype with characters controlled by the proposed architecture, while a second group will play the identical scenario with characters driven by more traditional, scripted Behavior Trees. Using post-session questionnaires and semi-structured interviews, I will analyze player perceptions of character intentionality, believability, and the legibility of their actions.

Broader Vision

My work is ultimately positioned to address a persistent challenge in interactive narrative: the tension between authorial control and character agency. This is often framed as a choice between "strong story" systems, like narrative planners (Ware and Young 2014; Ware and Siler 2021) that seek to add character motivation to an authored plot, and "strong autonomy" systems that prioritize emergent character behavior (Evans and Short 2014). However, a question arises from my proposed future work: if we allow characters to form their own supertasks, does this not ultimately destroy the author's narrative control? I argue that it does not, but rather shifts the focus of authorship from the direct scripting of actions to the design of the character's mind. This is a shift from the author as puppeteer to the author as gardener. Instead of pulling every string, the author designs the "seed": the *VivPersona* and *DELPEntity*, which encode the character's innate personality, core values, and understanding of the world.

The author also tends the "soil" by creating the narrative context and key events. In a theoretical recreation of *Macbeth*, an author would not explicitly assign the supertask *AvengeTheKing* to Macduff. Instead, they would design his *VivPersona* with core values of loyalty and justice. When confronted with the authored event of the king's murder, Macduff would autonomously form this supertask, as this goal is a coherent outgrowth of his fundamental character. Here, the designer's authored story exists as a powerful set of gravitational forces on the characters, rather than a rigid set of rails. The character tells the author's story not because it is a puppet, but because it is in its very nature to do so, allowing it to handle player deviations with an authenticity a scripted character never could.

Conclusion

To create believable characters driven by an explainable internal monologue, my doctoral research has produced two complementary prototypes: *Viv*, a complete architecture for goal attainment implemented in the game *Oops! All Bards*, and a foundational model for personality-driven goal formation within *Stay Thy Blade*. The remainder of my dissertation will synthesize these projects and conduct a multi-pronged evaluation of the final system. The ultimate contribution will be a validated framework that bridges the gap between agent autonomy and authorial intent, enabling the creation of "strong stories through strong characters."

Acknowledgements

I would like to acknowledge the support of the National Science Foundation under Grant No. IIS-2232066, as well as UC Davis and the Center for Artificial Intelligence and Computational Futures.

References

Agis, R. A.; Cohen, A.; and Martínez, D. C. 2016. Argumentative AI Director Using Defeasible Logic Programming. In *Computer Games*, 96–111. Cham: Springer International Publishing. ISBN 978-3-319-39402-2.

Alifieris, M.; Siolas, G.; and Stafylopatis, A.-G. 2023. Explainable Deep Reinforcement Learning: State of the Art and Challenges. *ACM Comput. Surv.*, 55(8).

Antoniou, G.; Billington, D.; Governatori, G.; Maher, M. J.; and Rock, A. 2000. A Family of Defeasible Reasoning Logics and Its Implementation. In *Proceedings of the 14th European Conference on Artificial Intelligence*, ECAI'00, 459–463. NLD: IOS Press.

Colledanchise, M.; and Ögren, P. 2018. *Behavior Trees in Robotics and AI: An Introduction*. Chapman 'I&' Hall/CRC Artificial Intelligence and Robotics Series. Boca Raton, FL: CRC Press, 1st edition. ISBN 9781138593732.

Evans, R.; and Short, E. 2014. Versu: A Simulationist Storytelling System. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(2): 113–130.

García, A. J.; and Simari, G. R. 2004. Defeasible logic programming: An argumentative approach. *Theory and practice of logic programming*, 4(1-2): 95–138.

Liao, F.; Gelardi, G.; Mitchell, K.; Sandhu, A.; and Mccoy, J. 2023. Sunset Valley: A Case Study in Computational Gossip. In *2023 IEEE Conference on Games (CoG)*. Boston, MA.

Mateas, M.; and Stern, A. 2002. A Behavior Language for Story-Based Believable Agents. *IEEE Intell. Syst.*, 17: 39–47.

Mateas, M.; and Stern, A. 2004. A Behavior Language: Joint Action and Behavioral Idioms. In Prendinger, H.; and Ishizuka, M., eds., *Life-Like Characters: Tools, Affective Functions, and Applications*, 135–161. Berlin, Heidelberg: Springer.

McCoy, J.; Treanor, M.; Samuel, B.; and Wardrip-Fruin, N. 2011. Comme il Faut: A System for Authoring Playable Social Models. In *Proceedings of the Seventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2011, October 10-14, 2011, Stanford, California, USA*, 158–163. AAAI Press.

McCrae, R. R.; and Costa, J. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2): 175–215.

Mitchell, K.; Pettijohn, C.; and McCoy, J. 2022. Never a Dull Moment: Believable Dynamic Character Beat Generation between Game World Events. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 18(1): 279–281.

Mitchell, K. D.; and McCoy, J. 2024. Exploring Stanislavskian Performance for Agent-based Nonplayer Characters through Defeasible Logic. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, IVA '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706257.

Orkin, J. 2006. Three States and a Plan: The A.I. of F.E.A.R. In *Game Developers Conference*.

Schell, J. 2019. *The Art of Game Design: A Book of Lenses*. CRC Press, third edition.

Shirvani, A. 2021. *Personality and Emotion for Virtual Characters in Strong-Story Narrative Planning*. Ph.D. thesis, University of Kentucky.

Stanislavski, C. 1989. *An actor prepares*. Routledge.

Stanislavski, C. 2013. *Building a character*. A&C Black.

Stanislavski, C.; and Hapgood, E. R. 2012. *Creating a role*. Routledge.

Treynor, N.; Mitchell, K.; Toothman, N.; Bloom, G.; Milburn, C.; Neff, M.; and McCoy, J. 2025. Modeling Conflict De-Escalation in Shakespeare through Hybrid NLP & Symbolic Approaches. In *2025 IEEE Conference on Games (CoG)*. To appear.

Ware, S. G.; and Siler, S. 2021. Sabre: A Narrative Planner Supporting Intention and Deep Theory of Mind. In *Proceedings of the Seventeenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 17, 99–106. AAAI Press.

Ware, S. G.; and Young, R. M. 2014. Glaive: A State-Space Narrative Planner Supporting Intentionality and Conflict. In *Proceedings of the Tenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 80–86. AAAI Press.

Świechowski, M.; Szymański, D.; and Mańdziuk, J. 2023. Learning Non-Differentiable Graphs of Utility AI. In *2023 IEEE Conference on Games (CoG)*, 1–8.