

Affect-Aware Agents for Emergent Social Conflict in Games

Weilun Deng

Worcester Polytechnic Institute, Worcester, USA
1335641879dd@gmail.com

Abstract. Affect-aware agents are one way to make social conflict between non-player characters in games feel more believable. Instead of relying only on fixed scripts, these agents maintain an internal model of emotion and social preferences that reacts to in-game events and to how other characters behave. When anger, gratitude, fear, or ambition change over time, the result may be alliance, betrayal, sacrifice, or a struggle for power between NPCs or between an NPC and the player. This paper presents a narrative literature review on how AI can recognize or simulate such emotions and social preferences to support emergent conflict. Prior work is organized into three strands: affective and motivation modeling for individual characters, social preferences and multi-agent learning for groups of agents, and drama or experience management systems that control the pacing and intensity of conflict events. For each strand, the review examines how state is represented, how conflict is triggered, and how player experience and safety are evaluated. On the basis of this review, a simple mechanism view is proposed that links emotional state and social values to typical conflict behaviors, and key design choices are identified that can make conflicts both understandable and controllable. The goal is to provide game designers and AI practitioners with practical ideas for using affect-aware agents when building games that rely on emergent social conflict.

Keywords: affect-aware agents, emergent social conflict, multi-agent reinforcement learning, drama management, game AI

1. Introduction

Digital games often include social worlds in which non-player characters (NPCs) cooperate with each other, compete for resources, and react to the player. Moments of betrayal, sacrifice, and power struggle are important in these worlds because they create tension and make relationships feel alive. In many games, however, such conflicts are controlled by fixed scripts or simple condition checks that fire at predefined points in the story. Scripted events can deliver a clear narrative, but they are usually not tied to an explicit model of what each NPC feels or what each NPC cares about. As a result, conflicts may appear arbitrary or repetitive, and they may not reflect what actually happened between the player and the NPCs in a particular playthrough.

This paper examines how AI can recognize or simulate emotion and social preferences in order to produce believable emergent conflicts among NPCs and between NPCs and the player. Affective and motivation modeling provides each character with an internal state that represents how it

appraises events, how it views other characters, and which goals it currently prioritizes [1-3]. When an NPC repeatedly experiences unfair treatment, its internal anger or resentment can increase and make punishment or withdrawal more likely. When an NPC receives steady help or recognition, its gratitude and attachment can grow, making protective or even self-sacrificial behavior easier to justify. In this way, emotions serve as a bridge between in-game events and visible social behavior such as betrayal, support, or reconciliation.

When agents share an environment and interact with one another, emotion alone does not fully explain how conflict develops. Research on social preferences and multi-agent reinforcement learning investigates agents whose reward depends not only on their own payoff but also on the payoff of others [4-6]. Different reward structures can make agents cooperate, exploit others, or retaliate, leading to different dynamics of cooperation, breakdown, and shifting alliances over time. At the same time, work on interactive narrative and drama or experience management examines how a system can coordinate the timing and intensity of events so that the narrative remains coherent and the player retains a sense of control [7-9]. These three areas— affective modeling, social-preference MARL, and drama management—are often explored separately, and limited research connects them with a focus on emergent social conflict in games.

The purpose of this paper is to organize and interpret existing work relevant to this topic. A narrative literature review is conducted to examine how current AI techniques model emotion and social preferences, how these mechanisms trigger conflict between NPCs, and how such conflicts are evaluated for believability and player safety. Prior work is grouped into three strands: (1) affective and motivation modeling for individual characters, (2) social preferences and multi-agent learning for groups of agents, and (3) drama or experience management systems that influence when conflicts appear and how strong they become.

This review is guided by three research questions:

RQ1: How can computational emotion models support believable behaviors such as betrayal, sacrifice, and power struggles?

RQ2: How do social preferences and multi-agent reinforcement learning contribute to emergent social behavior, and what controllability challenges arise for game design?

RQ3: How can drama or experience management systems determine when latent conflicts should surface while preserving player agency?

The rest of the paper is organized as follows. Section 2 describes the methodology of this narrative review. Section 3 reviews prior work across the three strands. Section 4 synthesizes these strands into a mechanism view. Section 5 discusses design implications and open problems. Section 6 concludes and outlines directions for future research on affect-aware agents for emergent social conflict in games.

2. Methodology

This paper adopts a narrative literature review approach to organize research on affect-aware agents and emergent social conflict in games. The objective is not to provide an exhaustive survey of all work on game AI or interactive narrative, but rather to collect representative approaches that illustrate how emotion and social preferences are modeled, how conflicts between non-player characters are triggered, and how these systems are evaluated. A narrative review is suitable here because relevant work spans multiple subfields—including affective computing, multi-agent reinforcement learning, and interactive narrative—which use different methods and vocabularies.

The primary sources for this review include the ACM Digital Library, IEEE Xplore, and major AI and game-AI conference proceedings such as AAI, AIIDE, and FDG. A small number of journal

articles from venues focusing on interactive narrative and virtual characters were also consulted. The time window for the search is approximately 2003 to 2025, covering both classic affective models that are still widely cited and recent systems that apply learning-based methods in interactive environments. All searches were conducted in English.

Search terms were derived from the three strands introduced in the literature review. For affective and motivation modeling, terms such as “affective computing,” “emotion model,” “appraisal,” “virtual character,” and “game NPC” were used. For social preferences and multi-agent learning, terms included “multi-agent reinforcement learning,” “social dilemma,” “social preferences,” and “inequity aversion,” combined with “game,” “environment,” or “agents.” For drama and experience management, terms such as “interactive narrative,” “drama manager,” “experience management,” and “social simulation” were used. When key papers were identified, their citation networks and reference lists were also examined to locate additional relevant work.

Screening was conducted in two stages. First, titles and abstracts were reviewed to exclude papers clearly outside the scope of this work, such as emotion recognition without interactive agents, reinforcement learning in purely physical control tasks, or narrative theory not tied to a running system. Second, the full text of remaining papers was examined, and only those meeting three criteria were retained: (1) they involve agents in an interactive digital environment (such as a game or simulation), (2) they describe an implemented or algorithmically detailed system rather than only a conceptual proposal, and (3) they relate directly to at least one of the three areas of interest (computational emotion modeling, social preferences in multi-agent learning, or drama and experience management that handles conflict events).

For each selected paper, a consistent set of features was extracted to support comparison. These features include the domain or game setting, how agent state is represented (e.g., emotion variables, social ties, learned value functions), how potential conflicts are triggered or resolved, and which evaluation measures are used. Comments on controllability, explainability, player experience, and safety were noted when available.

After extraction, the papers were grouped according to the three strands defined in Section 3 and re-examined in relation to the three research questions introduced in Section 1. The goal is not to create a comprehensive catalog, but to develop a structured synthesis that clarifies how existing AI techniques model emotion and social preferences to produce emergent social conflict between NPCs. Section 3 presents this review of prior work.

3. Literature review

3.1 Affective and motivation modeling

Affective and motivation modeling gives each non-player character an internal state that can support emergent social conflict. In many commercial games, emotions are represented as simple numbers or temporary status effects, which are increased or decreased by scripted events. In affective computing, by contrast, emotions are treated as the result of an appraisal process in which the agent evaluates what has happened, who is responsible, and how goals and relationships are affected [1]. Classic models such as the OCC model of emotion (a taxonomy that derives discrete emotions from appraisals of events, agents, and objects) and the EMA framework (Emotion and Adaptation, a continuous appraisal-and-coping model for dynamic environments) describe how events that help or hurt goals, actions that are judged as right or wrong, and changes in relationships can lead to anger, gratitude, guilt, fear, or other emotions [2,3]. These emotions can be stored as part of an NPC’s state, together with its goals and beliefs, instead of being hidden inside fixed story logic.

When an NPC has this kind of internal emotional state, the game can use it to drive more believable conflicts between characters. For example, if an NPC repeatedly appraises another character as blocking its goals or taking away resources, its internal anger or resentment value can gradually increase. Once this value passes a simple threshold, the NPC can start to select actions that distance it from the other character, reduce cooperation, or actively punish the perceived offender. From the player's point of view, this looks like a breakdown in a relationship or even a betrayal that follows from a history of unfair treatment. In the opposite direction, if an NPC receives frequent help, support, or recognition from another character, its gratitude and attachment values can grow over time. This makes it easier to justify later behaviors such as defending that character, sharing valuable items, or taking risks on their behalf. Designers can tune parameters such as how fast emotions increase, how quickly they decay, and which actions become available at which thresholds, so that conflict is tied to past interactions rather than to arbitrary triggers.

Affective and motivation modeling is closely related to the main topic of this paper, which is how AI can recognize or simulate emotion in order to support emergent conflict among NPCs. An emotion state that is based on appraisals gives a readable reason for why an NPC changed its behavior, which can be inspected through logs or visualization tools during development [1]. It also allows designers to connect abstract ideas such as loyalty, rivalry, or long-term resentment to concrete internal variables. At the same time, these models mostly explain why one character turns against another, and they tend to focus on small groups of well-defined agents. On their own, they say less about how larger groups form coalitions, enforce social norms, or collectively turn against a leader. For that reason, many researchers and designers consider affective modeling together with work on social preferences and multi-agent learning, which aim to explain conflict and cooperation at the group level [4-6].

3.2 Social preferences and multi-agent learning

Social preferences and multi-agent reinforcement learning (MARL) provide a way to study how conflict can emerge in a group of agents that share an environment. In many MARL setups, several agents collect resources, choose whether to share or withhold them, and sometimes have access to actions that punish or reward others [4]. Each agent learns a policy through repeated interaction, while the designer observes what kinds of social patterns appear. Some settings lead to stable cooperation, some produce free-riding and exploitation, and others show cycles in which groups form, break apart, and then re-form in new configurations [4]. This line of work focuses less on detailed internal emotions and more on how the reward structure and game rules shape long-run behavior.

To bring social preferences into this picture, researchers often change the reward that an agent receives so that it depends not only on its own payoff but also on the payoff of other agents. One common idea is inequity aversion, in which agents are penalized when their reward is much higher or much lower than that of their partners [5]. Another idea is social influence, where an agent receives additional reward when its actions change the behavior of others in a particular direction, for example toward cooperation [6]. In concrete implementations, these signals can take several forms. For instance, in a resource-harvesting gridworld, an agent may receive a small negative reward when choosing a special "punish" action that reduces another agent's return, or a positive shaping reward when its actions cause other agents to contribute more to a shared pool. In public-goods games, agents can be given a bonus when the group's total contribution exceeds a threshold, or a penalty when they take more than an equal share of a common resource. When such social signals are added to the learning objective, agents can learn to share resources, to punish agents that

take too much, or to respond strongly when they themselves are treated unfairly. In the same environment, different settings of the social preference parameters can therefore produce cooperation, escalating retaliation, sudden collapses of trust, or small stable coalitions [4-6].

For game designers who are interested in emergent social conflict between non-player characters, this research is both promising and challenging. It shows that complex group behavior can arise from relatively simple learning rules, without explicitly scripting who betrays whom and at what moment [4]. In principle, designers could train groups of NPCs with social preferences and then embed the resulting policies into a game, so that alliances, grudges, and shifts in power come from learned reactions to shared situations. At the same time, MARL systems are often sensitive to hyperparameters, random seeds, and details of the environment that are hard to control in a shipped product [5,6]. It can be difficult to guarantee that conflict will appear with the desired frequency and intensity, or to explain why a particular betrayal happened in one playthrough but not in another. For this reason, social-preference MARL is currently more useful as a way to explore possible dynamics and to inspire designs than as a black-box mechanism that directly drives all in-game conflict.

3.3 Drama / experience management and interactive narrative

Drama and experience management research asks how a system can control the timing and intensity of events in an interactive story while the player still makes meaningful choices. In traditional linear media, authors decide exactly when conflicts start, how they escalate, and when they are resolved. In games, players can move, act, and ignore hooks in ways that break a fixed plan. A drama manager is a layer that monitors the current state of the story and of the player, then selects events or suggestions that move the experience toward an author's high-level goals [7-9]. Classic work on guiding interactive drama [7], on interactive drama systems such as Façade [8], and on dynamic experience management in virtual worlds [9] all treat conflict as a central tool for shaping the player's experience.

Many drama management systems represent the story as a set of plot points, quests, or "beats" that can be triggered in different orders, subject to constraints [7,8]. The system tracks variables such as which characters have been met, which goals are active, and how much tension or uncertainty the player has experienced recently. When conditions are satisfied, the drama manager chooses one of several possible events, such as revealing new information, introducing a threat, or forcing a difficult choice. In a social game, this event may be a confrontation between NPCs, a disagreement with the player, or a change in who holds power in a group. Designers can mark some events as higher or lower in intensity, and can restrict how often high-intensity events occur, in order to avoid exhausting or confusing the player [9].

In the context of affect-aware agents and social preferences, drama and experience management can be seen as a coordinating layer. Instead of directly scripting that "NPC A betrays NPC B at time T," the system can monitor internal variables such as anger, fear, or inequity signals and wait until these values cross a range that makes betrayal plausible. At that moment, it can choose to foreground the conflict by scheduling a meeting, arranging that the relevant NPCs are present in the same location, or slightly adjusting the environment so that the confrontation becomes visible to the player. This allows emergent conflicts, which come from the agents' own states and rewards, to be aligned with pacing and presentation choices that support a clear story [7-9].

Drama management work also raises questions that are important for emergent social conflict. If a system pushes too hard, the player may feel that their actions do not matter and that the game is forcing a particular outcome. If it pushes too softly, important conflicts may never surface, or they may happen off-screen and fail to contribute to the experience. There is also a safety aspect: systems

that deliberately increase tension and interpersonal conflict must avoid creating situations that feel personally abusive or that repeatedly target vulnerable players [9]. For designers who want to combine affect-aware agents with emergent conflict, drama and experience management provide patterns for balancing author control, player agency, and emotional impact.

4. Synthesis

The literature reviewed in the previous sections suggests that affective modeling, social preferences in multi-agent learning, and drama or experience management can be viewed as three layers of a single mechanism for emergent social conflict in games. At the individual layer, affective and motivation models describe how non-player characters (NPCs) appraise events and update their internal emotional state [1-3]. At the group layer, social-preference multi-agent learning explains how agents respond to shared incentives and limited resources [4-6]. At the presentation layer, drama or experience management determines when and how conflicts that already exist in the agents' state are brought to the player's attention [7-9]. Together, these layers suggest a way to think about how AI can recognize or simulate emotion and social values so that conflicts appear to grow naturally out of play rather than out of arbitrary triggers.

From the perspective of individual characters, an NPC's emotional state is a compact summary of its recent history and priorities. Anger, fear, gratitude, and similar variables encode how the NPC has interpreted past events and how it currently evaluates other characters [2,3]. When these variables are linked to goals and action selection, they define a space of behaviors that includes loyalty, cautious cooperation, withdrawal, and open aggression. High anger combined with low trust in a particular partner increases the chance that the NPC will refuse help, hoard resources, or take a punishing action, whereas high gratitude and attachment toward a different partner make it more likely that the NPC will share rewards or accept personal risk. In this way, affective models help answer the first research question by showing how conflict-related behaviors can be traced back to interpretable internal states [1-3].

Social-preference multi-agent learning extends this picture from individuals to groups. When rewards depend partly on the payoff of others, agents develop preferences over distributions of outcomes, not just over their own score [4-6]. Inequity aversion encourages agents to avoid situations in which one member gains too much at the expense of the group, while social influence rewards agents that successfully shape others' behavior [5,6]. These mechanisms can lead to patterns such as stable cooperation, alternating exploitation, or the formation of small coalitions that protect their members and punish outsiders. In a game setting, such patterns correspond to factions that defend shared norms, groups that slowly turn against a selfish leader, or clusters of NPCs that coordinate to exclude another character. This line of work speaks to the second research question by showing that group-level conflict and alliance can emerge from learned responses to social rewards, but also by highlighting how sensitive these outcomes are to reward design and environment details [4-6].

Drama and experience management sit above these two layers and address a different part of the problem: when and how the player should see the conflicts that are latent in the agents' state. Even if anger, inequity signals, and coalition structures have developed in the background, the player may not notice them unless the game surfaces specific moments where they matter. A drama manager can monitor simple indicators, such as the distribution of emotional values, the recent frequency of hostile actions, or how long it has been since the last major conflict [7-9]. It can then decide whether to frame the next scene around a confrontation, a revelation, or a temporary resolution. In practice, this might mean arranging for two NPCs with high mutual resentment to appear in the same

location, or choosing a dialogue option that exposes a hidden grievance. Drama management therefore contributes to the third research question by offering tools that guide conflicts toward points in the story where they are both understandable and impactful, while still letting the underlying agents decide when they are ready to break [7-9].

In summary, affective modeling, social-preference learning, and drama or experience management each address a different aspect of emergent social conflict, and they can be combined into a coherent view of how AI supports such conflict in games. Affective models tie conflict to internal states that are understandable to designers and, in some cases, to players [1-3]. Social-preference learning shows how group dynamics can produce cooperation, exploitation, and shifting alliances in shared environments, but must be carefully controlled [4-6]. Drama and experience management connect these internal and group-level processes to the player-facing story [7-9]. Together, they suggest that believable emergent conflict requires not only expressive behavior from NPCs, but also clear internal state representations and higher-level control over when conflicts are allowed to surface.

5. Discussion

The synthesis in the previous section indicates that affective modeling, social-preference learning, and drama or experience management can be combined into a mechanism view of emergent social conflict in games. At the same time, the literature also highlights several tensions and open problems that are important for design. This section focuses on four themes: the balance between emergence and control, implications for practical game development, limitations of current approaches, and ethical and safety concerns.

A first theme is the trade-off between emergence and control. Affective models and social-preference multi-agent learning both introduce internal variables that evolve over time and can lead to conflicts without explicit scripting [1-6]. This supports replayability and can make NPC behavior feel less predictable in a positive way. However, the same mechanisms can produce outcomes that are hard to anticipate or to reproduce. If emotional thresholds are set too low, NPCs may break relationships too often and make the world feel unstable; if thresholds are set too high, conflicts may rarely surface and interactions may feel flat. In social-preference learning, small changes in reward parameters or environment details can flip a system from cooperation to constant retaliation [4-6]. For commercial games, where designers must deliver a specific level of difficulty and a stable experience, this uncertainty is a real obstacle.

For practical development, these observations suggest that emergent mechanisms are most useful when introduced in a controlled way rather than used as the sole driver of social behavior. One strategy is to use affective and social-preference models to generate candidate behaviors, but to filter or reshape them through higher-level rules or authoring tools [1-3]. Designers might specify ranges in which particular conflicts are allowed, define “safe” combinations of events, or add simple overrides that prevent extreme cases. Another strategy is to use emergent systems in limited parts of the game, such as side stories, small factions, or background simulations, while keeping the main narrative arc more tightly authored [7-9]. In both cases, the goal is to benefit from the variety and surprise that emergent conflict can offer without giving up the ability to tune pacing and difficulty.

A second theme is the need for tools that make these systems more transparent. Many studies assume that designers can inspect internal variables or replay logs, but do not provide concrete interfaces for doing so [1-3,4-6]. In practice, it is difficult to debug a system if an NPC’s betrayal can be traced only by reading raw numbers in a console. More usable tools could visualize emotion trajectories, resource exchanges, and social ties over time, and could highlight moments where

thresholds were crossed. Simple explanation generators could produce short summaries such as “NPC A refused to help NPC B because it felt repeatedly treated unfairly” based on internal state. Such tools would not only help with development, but could also support player-facing features that reveal parts of the NPCs’ reasoning in a diegetic way, for example through journals, rumors, or explicit dialogue [7-9].

A third theme concerns the limits of the current literature. Most systems are tested in relatively small environments, with a handful of agents and simplified social structures [4-6]. It is unclear how well these ideas scale to larger open worlds, where dozens of NPCs interact and where players can ignore or exploit systems in unexpected ways. Many studies also focus on short time spans, such as individual episodes or limited-length simulations, rather than on the long campaigns and repeated sessions that players experience in commercial games. In addition, work on affective modeling, social-preference learning, and drama management often uses different evaluation criteria, which makes direct comparison difficult [1-9]. There is a need for shared benchmarks and experimental setups that explicitly target social conflict and alliance in game-like environments, and that track both behavioral outcomes and player experience.

Finally, ethical and safety issues become more pressing as emergent social conflict is used in games. Systems that intentionally increase tension, jealousy, or distrust between characters might also create strong emotional reactions in players [7-9]. For some players, this is part of the appeal; for others, especially those who have experienced similar dynamics in real life, it may be distressing. Designers should consider including safety valves such as content settings that limit certain types of interpersonal conflict, opportunities for reconciliation, and clear feedback when the game is pushing events in a particular direction. There is also a risk that learned systems reproduce or amplify harmful stereotypes if they are trained on biased data or designed with unexamined assumptions about social roles [4-6]. Addressing these concerns requires careful testing with diverse players and an explicit design process that treats emotional impact as a primary consideration rather than an afterthought [1-3].

Overall, the discussion suggests that affect-aware agents for emergent social conflict are most promising when they are used as part of a layered design. Internal emotion and social-preference state can provide richer reasons for NPC behavior, while drama or experience management can keep conflicts aligned with pacing and player agency [1-9]. To turn this into a practical method for game development, future work needs to focus on transparent tools, shared evaluation environments, and design guidelines that balance emergence, control, and ethical responsibility.

6. Conclusion and future work

This paper reviewed how AI systems can model emotion, social preferences, and narrative control to support emergent social conflict among non-player characters. Affective and motivation models help explain individual NPC behavior by giving characters interpretable internal states. Social-preference learning shows how groups of agents can develop cooperation, exploitation, alliances, or rivalries based on shared environments. Drama and experience-management techniques provide higher-level control, determining when and how latent conflicts become visible to the player. Taken together, these approaches suggest that believable emergent conflict requires linking internal agent states with group dynamics and with presentation choices that fit the rhythm of play.

Overall, affect-aware agents are most effective when used as part of a layered design. Internal states offer reasons for NPC behavior, social-preference structures shape how conflicts escalate within groups, and drama management helps maintain coherence and player agency. When these

components reinforce one another, conflicts can appear natural, varied, and responsive to player influence.

Future work can proceed in several directions. One direction is the creation of shared benchmarks and evaluation environments that explicitly test how different systems model social conflict and alliance. A second direction concerns tools for designers, including clearer interfaces for inspecting agent states, explanations for NPC behavior, and visualization of emotion or relationship changes. A third direction is the study of player experience and ethics, especially in games where interpersonal conflict is intense or personally meaningful. Research on these topics will help guide the development of affect-aware systems that are both expressive and responsible, and that support richer interactions in future game worlds.

References

- [1] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.
- [2] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge, UK: Cambridge University Press, 1988.
- [3] J. Gratch and S. Marsella, “A domain-independent framework for modeling emotion,” *Cognitive Systems Research*, vol. 5, no. 4, pp. 269–306, 2004.
- [4] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, “Multi-agent reinforcement learning in sequential social dilemmas,” in *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2017, pp. 464–473.
- [5] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. A. Duéñez-Guzmán, A. García Castañeda, I. Dunning, T. Zhu, K. R. McKee, R. Koster, H. Roff, and T. Graepel, “Inequity aversion improves cooperation in intertemporal social dilemmas,” in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, 2018, pp. 3330–3340.
- [6] N. Jaques, A. Lazaridou, E. Hughes, Ç. Gülçehre, P. A. Ortega, D. J. Strouse, J. Z. Leibo, and N. de Freitas, “Social influence as intrinsic motivation for multi-agent deep reinforcement learning,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 3040–3049.
- [7] P. Weyhrauch, “Guiding interactive drama,” Ph.D. dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, Tech. Rep. CMU-CS-97-109, 1997.
- [8] M. Mateas and A. Stern, “Façade: An experiment in building a fully-realized interactive drama,” in *Game Developers Conference (GDC '03)*, San Jose, CA, USA, 2003.
- [9] M. O. Riedl, A. Stern, D. Dini, and J. Alderman, “Dynamic experience management in virtual worlds for entertainment, education, and training,” *International Transactions on Systems Science and Applications*, vol. 4, no. 2, pp. 23–42, 2008.