"Visible" Intelligence: Designing NPC Decision Explanation Interfaces Based on Web High-Fidelity Prototypes and Empirical Study of Player Trust

Aorui Huang
School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong, Shaanxi, 723000, China
huangaorui166@163.com

ABSTRACT

In today's video game world, the transformation from predictable state machine to deep reinforcement learning and generalized language model (LLM) has completely changed the mechanism of non-player role (NPC). However, although these generation agents show unprecedented intelligence, they also create a huge transparency vacuum. Players are increasingly talking to "black box" entities, and their reasoning is hidden, which breaks trust when there is AI illusion or unexpected tactical error. Based on the Situational Awareness Agent Transparency (SAT) model, we developed a highly accurate visual interpretation interface system for "translating" machine logic to players to dynamically calibrate trust. We use the Wizard of Oz technology to create a specially designed hybrid Web prototype to simulate the complex information of NPC. The behavioral data of 40 participants recorded with microsecond precision through the performance.now () API shows a very important result: providing complete transparency-especially showing the reasoning of artificial intelligence and its uncertainty-will not cause information overload. On the contrary, when AI fails, it greatly accelerates the intervention speed of players, improves the accuracy of decision-making, and greatly reduces the frustration of users. This study provides a solid framework and practical guide for developers who want to balance AI capabilities and human-centered interpretability in an interactive environment.

Keywords: Explainable AI; NPC Decision Transparency; Trust Calibration; SAT Model; Human-Computer Interaction.

1. INTRODUCTION

The era of predictable and scripted video game characters is rapidly ending. For decades, players have interacted with non-player characters (NPC), which work with simple finite state machines (FSM) or behavior trees. These early systems are easy to understand; Players can easily guess, "If I shoot the defender, he will hide". However, the integration of deep reinforcement learning and generative generalized language model (LLM) breaks this predictability. Modern NPCs can now adapt, formulate strategies and communicate at almost the level of human autonomy. However, this "intelligent explosion" has to pay a heavy price: with the increase of algorithm complexity, the logic of NPC's action becomes more and more invisible to human players [1].

When the player interacts with the "black box" agent, anything unexpected-such as suddenly shrinking in the battle he is about to win, or hallucinating the dialogue reaction-will immediately cause confusion. Without understanding why AI does this, players often imagine negative reasons. They may think that the game is not running well, the AI is "cheating", or the system is unfair. In the immersive world such as VR [2], this phenomenon is more intense, and the unexpected behavior of NPC can completely break the sense of immersion [3]. Therefore, the challenge for modern game designers is not only to make NPCs smarter, but also to make them more understandable. We need to go beyond blind trust and enter a "trust calibration" state, where players know exactly when they can rely on AI and when the trust is too low to trust [4, 5].

In order to cope with this transparency crisis, our research adapts to the situation-based proxy transparency (SAT) model, which was originally developed for high-risk military environments and automated vehicles [6]. We designed a high-fidelity UI visual component library to externalize the internal reasoning of the biological NPC. Using the design-oriented Web prototype method, we try to quantify the precise influence of transparency level on players' behavior and psychological burden. Our main goal is to determine whether exposing the "mathematics behind madness"-including the reasoning weight of artificial intelligence and its survival prediction-can help players maintain a stable cooperative relationship with their digital teammates, even if the teammates make catastrophic mistakes.

2. LITERATURE REVIEW

2.1 The Rise of Explainable AI (XAI) in Gaming

The field of interpretable artificial intelligence (XAI) emerged to cope with the lack of transparency of deep learning models. Traditionally, XAI tools such as salinity maps or locally interpretable model diagnostic interpretation (LIME) are designed to help data scientists debug. However, the transformation of teamwork between human and artificial intelligence needs "people-centered XAI", which transforms the complex weight of the model into an intuitive metaphor of non-experts. This is especially important in the field of games.

Morrison et al. (2023) showed through their "Eye into AI" project that the interactive platform can better evaluate how humans really understand the interpretation of AI than the static survey [7]. The game provides a unique environment, and users' understanding of AI has a direct and measurable impact on their behavior and success. Recent studies have also pointed out that the presentation of these explanations-whether they seem confident or cautious-has greatly changed the way humans rely on the results produced by LLM [8].

2.2 The Three Layers of the SAT Model

A big problem of transparency is "information dumping", when there is too much data and users are lost. The SAT model proposed by Chen et al. (2018) solves this problem by dividing the information into three layers [6]. The first level (perception) only tells the current state of human agents and their near-term goals. Level 2 (understanding) provides "why" by showing the reasoning, constraints and priorities behind it. Finally, the third level (projection) allows agents to predict what will happen, and most importantly, measure and say it when uncertain; Although simple symbols of level 1 are common in modern games (such as "in trouble" icon), to truly calibrate trust, artificial intelligence must honestly treat its limitations, which is only possible at level 3[6].

2.3 Automation Bias, Trust Junk, and Over-reliance

The ultimate goal of NPC transparency is not to make players trust AI more, but to make players trust AI correctly. When the user follows the advice of the machine without thinking, automatic deviation will occur, even if the machine is obviously wrong [9]. This is usually caused by the "concealment" that the machine may fail. As Lee and See (2004) said, trust calibration is a dynamic process, and our dependence must match the changing ability of the system [4]. Designers must be careful not to create "Trust Junk"-a flashing interface element that artificially increases trust without providing useful reasoning [10]. When a system does not admit its doubts, it may lead to dangerous "over-dependence" [11], and users stop checking the work of the machine, thinking that it will never go wrong. Therefore, an effective XAI must act as a "trust signal" so that human beings can recalibrate their trust in real time [12].

3. METHODOLOGY

3.1 Design-Driven Web High-Fidelity Prototyping

A common trap in HCI research is "minimum true deviation", in which participants give poor marks to the system just because the graphics look fuzzy. In order to ensure ecological effectiveness, we avoid wireframes, and instead build a high-fidelity tactical game interface using Figma and advanced CSS/HTML5. The visual style follows Cyberpunk's aesthetics, and the deep space background (#0a0818) is compared with the bright blue of standard HUD element (#66fcf1). We carefully designed the SAT interface layer to make it both immersive and informative.

- Perception (Level 1): Intention badge, which changes color according to NPC status. For example, cyan means "Tactical Cover" and flashing red means "Aggressive Charge".
- Understanding (Level 2): Logical panel with drop-down menu. There are bars showing the importance of different things to decision-making, such as "Enemy Fire Density" and "Ally Proximity".
- Projection (Level 3): Probability dashboard, which shows the percentage of "survival success rate" and whether it is safe or not, such as "best" or "fatal risk".

3.2 The "Wizard of Oz" Simulation Approach

In order to separate the interface effect from the characteristics of a specific AI model, we use the Wizard of Oz (WoZ) method. Participants think they are interacting with a complex, self-learning NPC. In fact, the behavior of NPC and the corresponding UI data are secretly controlled by the hidden experimenter through the password-protected backend [13].

This "puppet master" setting enables us to inject a specific "AI illusion"-when NPC inexplicably decides to give up its cover and load it into some kind of death. This controllable failure is the only way to test whether players really understand the SAT interpretation.

3.3 High-Precision Timing with performance.now()

Measuring human reaction time (RT) in browser experiment needs scientific accuracy. The JavaScript standard Date.now () is unreliable because it is synchronized with the system clock, making it sensitive to clock and NTP updates. According to the strict standard of bridges and his colleagues (2020), we use the performance.now() API [14]. The accuracy of this monotonous clock is less than milliseconds (microseconds), and it is not affected by the time change of the system, ensuring that the intervention speed we recorded is as accurate as that of the special psychology laboratory. This kind of accuracy is very important to identify the small differences in cognitive processing speed triggered by different transparency levels.

3.4 Experimental Design and Procedure

The experimental design between-subjects was used in this study. Forty participants were recruited for the experiment. Established the inclusion criteria of at least 5 hours of game experience per week. This specific threshold is based on the literature of Cognitive Science [15], which classifies individuals who play fast-paced video games for at least 5 hours a week as action video game players. This regular exposure ensures that participants have developed the necessary visual attention and spatial resolution to quickly process the complex and dynamic information on the head-mounted display (HUD), thus eliminating the basic confusion with the interface as a chaotic variable. These participants were randomly assigned to one of four conditions (n=10 in each group):

- Group 0 (Control): Complete black box (no XAI interface).
- Group 1 (SAT Level 1): Only intent tags displayed.
- Group 2 (SAT Level 2): Intent tags and reasoning weights displayed.
- Group 3 (SAT Level 3): Full transparency, including uncertainty projections.

During the simulation, WoZ operators will trigger NPC-specific tactical errors. In order to strictly measure the behavior, we have established a strict operational definition for the player's response to these NPC errors: (1) When the player accurately identifies the fatal error and presses the overwrite button in the allowed response window (for example, 3,000 milliseconds), the NPC will perform the action in an irreversible way. (2) If the player does not provide any input within this critical value of 3000 milliseconds, a timeout will occur, which will lead to wrong operation. (3) Failure includes timeout and false alarm, in which the player improperly triggers rewriting in the normal and safe operation of NPC.

The system will record the reaction time of the participants to cover the NPC. After completing the task, participants will fill in S-TIAS [16] to measure immediate confidence, and NASA-TLX [17] to evaluate cognitive load and depression level.

4. RESULTS AND DISCUSSION

4.1 Behavioral Performance: Accuracy and Reaction Time

The accuracy of decision-making-the rate at which players try to catch NPC errors-shows a clear linear trend. In group 0 (control), the success rate is only 40%. It rose to 80% in the first group and reached a peak of 90% in the second group. In the third group, the ratio remained at 80%. This shows that even the basic intention visualization (Level 1) is a powerful tool to break the bias of automation.

Reaction time (RT) data gives the most interesting results. In group 0, the average reaction time of successful overuse was "$M = 1315.00$" ms ("$SD = 210.4$"). Groups 1 and 2 have similar performance ("$M = 1289.40$" ms and "$M = 1311.90$" ms, respectively). However, the participants in the third group (level 3) are much faster, with an average reaction time of "$M = 1064.40$" ms ("$SD = 145.2$"). One-way ANOVA confirmed that this difference was significant, with "$f(3,36) = 4.21$", "$p = 0.015$" and $\eta_p^2 = 0.26$. This result shows that the idea of "information overload" is incorrect here. People may think that the complex survival probability of level 3 will slow down the speed of users. On the contrary, a clear message "confidence warning" acts as a psychological shortcut. By admitting their doubts, NPC actually eliminates the need for players to analyze the situation from the beginning, which makes the response speed almost 250 milliseconds faster.

4.2 Trust Calibration and the Prevention of Collapse

The confidence scores measured from 1 to 7 by using three items of S-TIAS showed a catastrophic breakthrough in the control group ("$M = 2.10$", "$SD = 0.82$"). Without any explanation, players think that NPC's mistake is a sign that the game is broken. However, the third group maintained a very high trust score of "$M = 6.00$" ("$SD = 0.64$"). This perfectly illustrates the "trust calibration" [12]. Since the NPC in Group 3 uses its Level 3 interface to show that it is not very safe before failure, players regard the error as a predictable marginal situation, not an intelligence failure. By demonstrating its vulnerability, artificial intelligence actually strengthens the long-term partnership between human beings and artificial intelligence.

4.3 Cognitive Workload and the Paradox of Information

The results of NASA-TLX reveal a paradox: the biggest mental load is the group with the least information. Group 0 reported a frustration score of "$M = 85.10$", while Group 3 reported almost zero frustration ("$m = 6.80$", "$p < 0.001$"). Although the third group has more UI elements to monitor, their psychological needs score is the lowest ("$M = 26.80$"). This proves that the cognitive effort required to guess the machine's intention is far greater than the effort required to read and explain. Transparency turns artificial intelligence into a predictable teammate and a highly stressful monitoring task into a low-stress collaborative experience.

5. DESIGN IMPLICATIONS FOR DEVELOPERS

According to our empirical research results, we provide three main guidelines for developers to integrate generative AI NPC in an interactive digital environment:

- In fast interaction, players don't need to know every variable in the neural network. They need a clear, very obvious "trust signal". When AI is unsafe, the user interface must explicitly require human intervention.
- Explain logic, not mathematics: Level 2 reasoning should use relative weights (bars or colors) instead of raw numerical data. Causal logic ("Why am I hiding") is more useful to players than coordinate lists.
- Admit vulnerabilities early: In order to prevent trust from collapsing after illusion, NPC must admit its doubts before the action is fully implemented, and preventive transparency is the only shield to prevent completely giving up generating agents after a mistake.

6. CONCLUSION

This study shows that the SAT model is an important basis for the next generation of transparent NPC. By stopping using the "black box" proxy and adopting high-fidelity visual interpretation, we can achieve real trust calibration. Our experimental results on 40 players show that displaying "why" and "doubt" behind AI behavior will not make human users overworked. On the contrary, it greatly speeds up the response time, almost eliminates the user's frustration, and prevents the trust from collapsing when an error occurs. For game developers who use LLM, the way forward is clear: be honest with players. Showing NPC's reasoning and admitting its doubts is a promising way to create a reliable digital trailer that players can really rely on.

REFERENCES

[1] Yin, M., Wang, E., Ng, C., & Xiao, R. (2024, May). Lies, deceit, and hallucinations: Player perception and expectations regarding trust and deception in games. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-15).
[2] Bazzaz, M., & Cooper, S. (2026). Playing the Imitation Game: How Perceived Generated Content Shapes Player Experience. arXiv preprint arXiv:2602.14254.
[3] Korkiakoski, M., Sheikhi, S., Nyman, J., Saariniemi, J., Tapio, K., & Kostakos, P. (2025). An Empirical Evaluation of AI-Powered Non-Player Characters' Perceived Realism and Performance in Virtual Reality Environments. arXiv preprint arXiv:2507.10469.
[4] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.
[5] Romeo, G., & Conti, D. (2026). Exploring automation bias in human-AI collaboration: a review and implications for

explainable AI. Ai & Society, 41(1), 259-278.
[6] Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical issues in ergonomics science, 19(3), 259-282.
[7] Morrison, K., Jain, M., Hammer, J., & Perer, A. (2023). Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game with a Purpose. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), 1-22.
[8] Kim, S. S., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025, April). Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-19).
[9] Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., & Ma, X. (2025, April). Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-23).
[10] Wall, E., Matzen, L., El-Assady, M., Masters, P., Hosseinpour, H., Endert, A., ... & Padilla, L. (2024, April). Trust junk and evil knobs: Calibrating trust in AI visualization. In 2024 IEEE 17th Pacific visualization conference (Pacificvis) (pp. 22-31). IEEE.
[11] Tatasciore, M., & Loft, S. (2025). Calibrating reliance on automated advice: transparency and trust calibration feedback. International Journal of Human-Computer Interaction, 41(23), 14723-14733.
[12] Li, Z., & Steyvers, M. (2026). Learning to Trust: How Humans Mentally Recalibrate AI Confidence Signals. arXiv preprint arXiv:2603.22634.
[13] Porcheron, M., Fischer, J. E., & Reeves, S. (2021). Pulling back the curtain on the wizards of Oz. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW3), 1-22.
[14] Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. PeerJ, 8, e9414.
[15] Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. Psychological science, 18(1), 88-94.
[16] McGrath, M. J., Lack, O., Tisch, J., & Duenser, A. (2025). Measuring trust in artificial intelligence: Validation of an established scale and its short form. Frontiers in Artificial Intelligence, 8, 1582880.
[17] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology (Vol. 52, pp. 139-183). North-holland.