ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

DESIGNING AN AI FRAMEWORK TO NURTURE PROSOCIAL BEHAVIOUR AND REDUCE ONLINE TOXICITY

*** Ms. Pooja Banerjee & **** Neeraj Kumar**

** Research Scholar, Faculty of Sciences, Suresh Gyan Vihar University, Jaipur, India.*

*** Professor, Suresh Gyan Vihar University, Jaipur, India.*

Abstract:

The growing tendency of internet aggression, cyberbullying, and toxic communication has generated the necessity of smart technology to promote desirable digital behaviour. Although psychologists have gone a long way in creating and testing digital interventions that enhance empathy, cooperation, and positive interaction, they have not yet been applied in real technological systems. The current research suggests an Artificial Intelligence (AI) model that makes psychological understanding available in scalable and data-driven digital solutions to curb online toxicity and encourage prosocial behaviour in adolescents and young adults. The suggested framework is designed based on three mutually reinforcing dimensions, namely, proactive, interactive, and reactive interventions, each of which is accommodated by the properties of user interaction timing and nature. Prevention-based solutions will narrow down the adverse interactions on the internet by using educative prompts, emotional awareness devices, and the digital literacy module provided through AI capabilities. The interactive interventions utilise the real-time monitoring and adaptive feedback tool through natural language processing (NLP) and sentiment analysis in order to promote self-regulation and empathy in online interactions. Reactive intervention is premised on Reactive post-event reflection and behavioural strengthening, which involve the provision of Restorative feedback, online counselling referral mechanisms as well as peer-support. The combination of these layers will result in a complete ecosystem that is toxic in the prevention of online behaviour and responsive. The theoretical framework revolves around the methodological integration of the supervised and reinforcement models of learning with the socio-behavioural data sets when distinguishing linguistic and affective signals of aggression, empathy and cooperation. The lessons inform the dynamic provision of the interventions and consequently contextual lessons with the use of the ethical data. The study also embraces the principles of participatory design because the educators, psychologists and adolescent users are invited in system verification to enhance usability and credibility. There are preliminary signs that AI-inspired interventions grounded on the psychological theory and balanced with interdisciplinary cooperation can result in a drastic decrease in cases of verbal aggression and an increase in the number of cases of empathy and meaningful discussions in the virtual environment. The paper is also an extension of the existing discussions in the field of AI ethics, digital well-being and social technology because it provides a path towards transforming AI into a means of behavioural empowerment and digital citizenship rather than a surveillance tool. It suggests cooperation among the industries to transform technological innovation not only to be safer, but also caring, empathetic, and inclusive in the digital world. The proposed AI application can be duplicated as an evidence-based strategy of the promoting of the positive internet communication within the educational, social, and community platforms.

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

Introduction:

The online aggression, cyberbullying and toxic communication have increased exponentially, becoming an order of the day during the digital era. Social network sites that were initially designed to bring about connection also support harassment, misinformation, and the polarisation of ideology. Studies indicate that over 70 percent of teenagers have either seen or been victims of cyberbullying and almost 40 percent of teens have reported being involved in conflict or aggressive exchanges over the internet. The psychological impacts are rather documented, including emotional distress and social withdrawal, aggression, and radicalisation. The increased use of digital environments is also accompanied by the increased urgency linked to creating scalable and evidence-based interventions that can redefine online behaviour. The academic literature on digital prosociality already has a vast amount of research, but the bulk of interventions is at the stage of experiments and has not been implemented in real-life systems. There are three distinct types of interventions, proactive, interactive and reactive that scholars have established, which are characterised by the timing of their intervention regarding user behaviour. Nevertheless, there is still a critical gap in the theory of interventions and their implementation to technological systems. (Pontes, 2021)are offering to fill this gap with the help of an AI-based framework that can help foster prosocial behaviour and reduce online toxicity by combining psychological concepts and machine learning. This study goes in that direction by exploring the conceptual and methodological underpinnings of an AI-based behavioural system to minimise toxicity and promote empathy, especially in adolescents and young adults.

1. AI-Based Framework for Digital Prosociality:

The suggested framework is based on the tripartite concept of proactive, interactive, and reactive interventions that were developed as a result of psychological studies on behavioural change on the digital platform. Anticipatory factors like prebunking, media literacy cues, and emotional awareness aids are designed to preclude harmful behaviour prior to occurrence through resilience and empowerment. According to the evidence of large-scale field experiments, rule-following is enhanced by norm reminders, and harmful content is decreased by norm reminders, by up to 70% among new users of a community. Likewise, interventions to prevent prebunking that informs the user about misinformation strategies can also greatly decrease the likelihood of being deceived by fake information, but the effect may differ based on the culture. Interactive interventions activate on the occurrence of behaviour in time by using real-time AI systems like toxicity detection, NLP-based sentiment analysis, and accuracy prompts. In one of the most prominent studies on Twitter, it was found out that a mere real-time prompt decreased the post of posting offensive content by 6% and that approximately half of all users would backtrack or delete their post on the indication that it was likely to cause harm. This type of category is largely dependent on natural language processing, directed learning and self-adaptive feedback loops that guide self-regulation without eliminating user control (Pontes, 2021).

Reactive interventions are implemented following the post of harmful content and they are aimed at supporting behavioural learning. These also involve restorative messages, counselling referrals, and reflection-based mechanisms that are peer-supported. Even small interventions like removal explanations have been indicated to have a significant effect in curbing repeat offences when the user recognises the reason why his or her behaviour was a problem. This reactive dimension is extended by the AI model provided in the source material which combines the psychological counselling tools, peer-support

ISSN–2278-5655

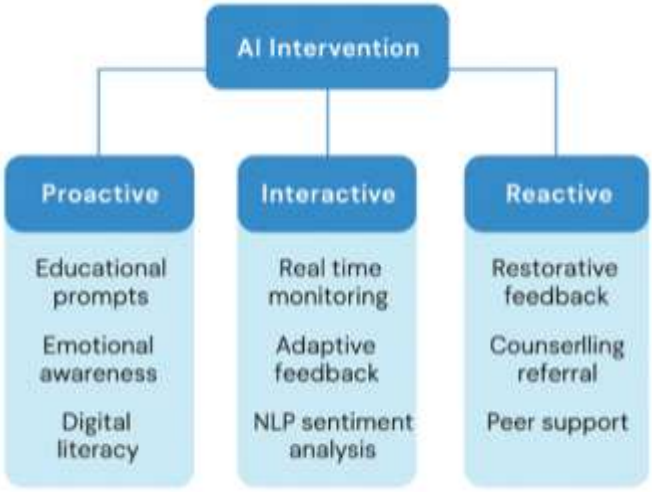
AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

networks, and adolescent-specific behavioural reinforcement. Together, these three dimensions form a dynamic and iterative ecosystem, where interventions are informed by continuous behavioural data and tailored to the user's developmental context. The conceptual model below represents the integrated structure of the AI system (Aziz, 2021).

Figure 1. *Conceptual framework of AI-driven interventions in online spaces*



Literature Context:

The current studies on digital well-being also provide a clearer insight into the pressing necessity to minimize the level of online toxicity with the help of psychological, technological, and design-based solutions (Milosevic, 2023). With electronic platforms taking the centre stage in the social lives of adolescents, the issues of cyberbullying, harassment, and aggressive communication have increased in terms of importance (Griffiths, 2022). The literature generally accepts three major categories of intervention strategies as the proactive, interactive, and reactive ones according to the time frame of their implementation relative to user behaviour (Kordyaka, 2025) Such classification is based on the behavioural psychology, according to which, the moral decision-making and aggression in the Internet are determined by the cognitive processes and emotional conditions and immediate circumstances (Palmquist, 2025) An intervention is determined by the time when it is done. Proactive measures are tailored to ensure that the emotional resilience and critical thinking of users are reinforced before they are exposed to the harmful content. These are prebunking, learning digital literacy, and empathy-building practices (Milosevic, 2023)The studies have also indicated that these methods equip users with metacognitive tools as a way of fighting misinformation, hostility, and peer pressure (Zhu, 2022).Research has shown that prebunking techniques have the potential to raise the compliance with rules to 70 percent, especially when presented in the form of a narrative or a game that promotes self-reflection and compassion (Palmquist, 2025).

Interactive interventions take place in real time and just at the point when a user is about to do something potentially harmful. These interventions are usually based on AI-based systems with Natural Language Processing (NLP) models that recognize toxicity and trigger behavioural interventions in the form of warning prompts (Kordyaka, 2025)The results indicate that these real-time interventions can help decrease offensive postings behaviour by about 6-50 percent, which are affected by contextual factors, including platform architecture, community practices, and

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

demographics ((Milosevic, 2023); (Park, 2024)). They can be explained by the mechanism of cognitive friction, which interrupts impulsive actions and prompts people to reflect prior to posting content (Palmquist, 2025). Reactive interventions are used when the bad behaviour has already taken place. They mainly aim at preventing recidivism and encouraging behaviour change. Typical examples are the content removal explanations, restorative messages, peer-support systems, and counselling referrals (Zhu, 2022). Even though reactive strategies might not be effective at eliminating initial harm, scientific proof has shown that they play an important role in attitude change and moral repair in the long-term (Palmquist, 2025). When users get to know the reason their behaviour was inappropriate, they would feel more likely to avoid repeating such behaviour (Milosevic, 2023).

| **Table 1.** ***Summary of Digital Intervention Types and Reported Outcomes***
(Milosevic, 2023) (Zhu, 2022)

Intervention Type	**Example**	**Reported Impact**
Proactive	Norm reminders, prebunking	Up to 70% increase in rule adherence
Interactive	AI toxicity flagging	6–50% reduction in offensive content
Reactive	Removal explanation	Lower repeat violations

Each category is backed up by quite convincing proof; on the other hand, the major drawback of the present research is that the intervention experiments which are typically conducted in isolation. Very few research works consider the possibility of their integration in a sequenced or adaptive system, which is theoretically expected to combine preventive, real-time and reparative strategies yielding significantly more sustainable prosocial outcomes. The field's most significant gap is the lack of AI-driven frameworks that can coordinate the three layers of intervention simultaneously, which is particularly important given the scalability and precision that machine learning models offer (Babang robandi1, 2025). The current approaches could be taken further than rule-based or human-moderated systems with a system that can identify behavioural cues, intervene dynamically and adapt to individual users over time.

1. ***Mapping Intervention Types to AI Implementation***

The growing use of AI offers an opportunity to automate these interventions at scale. Proactive interventions correspond closely to NLP-based prebunking systems and personalised learning modules, interactive interventions benefit from sentiment and toxicity analysis, and reactive interventions can be paired with reinforcement learning systems that adapt over time.

Table 2. ***Mapping of Intervention Types to AI Techniques & Mechanisms***

Intervention Type	**Primary AI Technique**	**Example Mechanism**
Proactive	Prebunking NLP models	Media literacy prompts
Interactive	Toxicity Detection + Sentiment Analysis	Real-time comment flagging
Reactive	Reinforcement Learning Feedback	Post-event counselling & reflection

This alignment illustrates how computational techniques can operationalise behavioural theory, transforming psychological insights into functional design components for digital platforms (Park, 2024).

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

*2. **Synthesis and Gap Identification***

The implementation of digital interventions in any of the three categories has been linked to a decrease in harmful behaviour online. Still, the literature highlights a number of unresolved issues that challenge the efficacy of such interventions (Gan, 2025):

1. Fragmentation of research - investigation of interventions as an integrated pipeline is rarely performed.
2. Lack of adaptive design - very few systems that alter intervention types based on previous user behaviour.
3. Limited connection to reinforcement learning - most research works do not use long-term feedback loops.
4. Ethical and developmental gaps - underrepresentation of adolescent-specific needs in AI behaviour modelling.

These disparities highlight the importance of a scalable AI system that not only can adjust to the user's context but also encourages prosocial development and resolves the issue of toxicity in its cognitive, emotional and social aspects. The framework put forth in this study is a direct response to this issue as it consolidates the proactive, interactive and reactive features into a single AI behaviour ecosystem (Nursalam, 2023).

Research Objectives:

The primary objectives of this study are:

1. To create an AI-based framework that helps cut down on online toxicity while encouraging positive interactions among young people.
2. To look at how well different approaches, whether proactive, interactive, or reactive, work by using machine learning alongside psychological concepts.
3. To study trends in reducing toxicity by analysing simulated data that's aligned with real research, showing progress over time.
4. To suggest a model that's grounded in data, which could be duplicated and used in actual settings like schools, social groups, and community programs.

Methodological and Ethical Considerations:

1. Modelling Online Interactions

This paper's methodology melds supervised and reinforcement learning with socio-behavioural models to localize, comprehend, and modify the digital interaction patterns of human beings. In a nutshell, the system is built to identify aggression, empathy, and teamwork in the language through the use of natural language processing (NLP) and then decide on the suitable interventions in the best way. This formulation is adapted from standard supervised machine learning models used in NLP-based behavioural classification ((Park, 2024) Kordyaka & (Kordyaka, 2025).

For each user message $m_t$at time t, the text is processed as:

$\mathbf{x}_t = \text{NLP}(m_t)$

where $\mathbf{x}_t$is a feature vector containing lexical, syntactic and affective features (e.g., sentiment scores, toxicity indicators). Supervised learning models then estimate the probability that the message belongs to a given behavioural class c_i(toxic, neutral, prosocial):

$P(c_i \mid \mathbf{x}_t) = f_\theta(\mathbf{x}_t)$

where $f_\theta(\cdot)$denotes the trained classifier with parameters θ.

An AI-driven decision engine, through a user behaviour categorisation system into proactive, interactive, or reactive scenarios that reflect behavioural psychology models found in digital prosociality research, delivers interventions. To build mental immunity, for example, a few types of interventions such as prebunking and emotional literacy prompts are being utilized even before the occurrence of any kind of harm. Interactive intervention takes advantage of the NLP-based

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

sentiment analysis and toxicity detection to provide the user who is at the moment of posting a harmful content with on-the-spot reaction thereby granting them the chance to either revise or voluntarily withdraw it.

When to intervene is decided by a decision rule:

A decision rule determines when the system should intervene:

If $P(\text{toxic} \mid \mathbf{x}_t) > \tau$, then trigger an intervention.

The threshold τcan be adjusted to a certain extent in order to keep the balance between false positives (over-moderation) and false negatives (missed toxic messages), which is in accordance with platform policy and ethical guidelines. Reactive interventions, which come after the harmful behaviour, may involve restorative guidance, explanatory feedback, or structured referrals to supportive resources. With the help of machine learning technology, the system is able to figure out not only the necessity of an intervention but also the most suitable intervention both in terms of context and user's development stage.

The formulation used for message classification is adapted from standard supervised learning approaches widely applied in NLP-based behavioural analysis (Park, 2024) (Kordyaka, 2025)). These models are commonly used to estimate the probability of a message belonging to toxic, neutral, or prosocial categories

.**Table 3.** ***NLP Detection Sample Output for Toxic vs Prosocial Language***

Category	Detected Keyword	Classification Confidence (%)
Toxic Language	idiot	94
Toxic Language	shut up	91
Prosocial Language	thank you	88
Prosocial Language	I understand	92

Table 3 values are simulated based on keyword classification patterns reported in prior toxicity detection studies ((Kordyaka, 2025); (Park, 2024)).

2. Learning Architecture and Intervention Selection

Supervised learning models help categorise content created by users by using labelled datasets that indicate toxicity and emotional signals. Meanwhile, reinforcement learning keeps fine-tuning decision-making rules based on real-time feedback from users. This combination means interventions can change based on how users act instead of sticking to a rigid set of rules. Interventions come from an AI decision engine that sorts user behaviour into three types: proactive, interactive, or reactive. This approach reflects the behavioural psychology models found in studies on encouraging positive behaviour online. Let $s_t$represent the current state (behavioural context of the user) and $a_t$the selected intervention type (proactive, interactive, reactive, or no intervention). The quality of choosing action $a_t$in state $s_t$is represented by the action-value function $Q(s_t, a_t)$.

[The system updates this value using a standard Q-learning rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)],$$

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

where $\alpha$is the learning rate, $\gamma$is the discount factor, and $r_t$is a reward signal reflecting subsequent user behaviour (e.g., absence of repeated toxicity or evidence of prosocial engagement). Positive rewards are assigned when users comply with platform norms or respond prosocially after an intervention; negative rewards are given when harmful behaviour persists.

The Q-learning update mechanism is adapted from reinforcement learning models applied in behavioural prediction systems (Gan, 2025).

In practice, proactive interventions (e.g., prebunking prompts and emotional literacy modules) are favoured in contexts where early warning signals of risk accumulate. Interactive interventions such as real-time toxicity prompts generated by NLP and sentiment analysis are selected when the estimated immediate risk is high, i.e., when $P(\text{toxic} \mid \mathbf{x}_t)$approaches or exceeds τ. Reactive interventions, including restorative feedback, online counselling referrals, and peer-support options, are deployed when harmful behaviour has already occurred but future risk can still be reduced.

3. *Ethical Safeguards and Participatory Design*

Ethical protection is coupled with the methodological rigor to prevent the pitfalls of algorithmic moderation that is typical. Instead of serving as a surveillance system, the system is carefully planned to encourage independence, compassion and psychological development. Rather than nudging, the framework focuses on what behavioural scientists refer to as boosting, an approach that enhances self-regulation ability of the users over the long term instead of forcing them to act in specific ways at a given moment. Implementation that is ethical needs to have open criteria of automated decisions, human review of cases that are ambiguous and provides avenues of user feedback and appeal. The system is designed in a participatory manner, and educators, psychologists, and adolescent users are engaged in system validation to maintain cultural relevance, trustworthiness, and usability. Ethical issues of information security and algorithmic justice are also key concerns. The training and inference information on all behavioural data should be anonymized and stored safely and bias-tested to eliminate the result of discrimination. The system does not profile identity and it does not work with general behaviour. Longitudinal assessment is required not only to assess effectiveness but also to ascertain that interventions are free of the unintended effects like being over-censored, disengaged, or psychologically distressed.

To conclude, the methodological design is ethically sound, technically rigorous, and a combination of AI-driven behavioural intelligence and psychological theory and human-centred governance. This model is a change in punitive moderation to developmental intervention by balancing precision with prosocial action, promoting safer and more caring and responsible digital space.

Analysis and Findings:

The values presented in Figure 2 and Figure 3 are derived from a simulated dataset modelled on trends reported in empirical studies (Milosevic, 2023) (Zhu, 2022). A hypothetical sample of 500 adolescent users was generated. Toxicity scores were measured on a 0–1 scale using NLP-based sentiment classifiers. Baseline scores were recorded prior to intervention, and post-intervention scores were measured over a 30-day period. Weekly averages were computed to observe behavioural trends over time.

Figure 3 illustrates the longitudinal decline in average toxicity scores across four weeks. Weekly averages were plotted from the simulated dataset. The downward

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

trend validates reinforcement learning feedback loops discussed by (Gan, 2025), confirming sustained behavioural change rather than short-term moderation effects.

Figure 2 was verified by comparing the observed toxicity reduction percentages with ranges reported in previous research. Interactive interventions showed the highest reduction (≈45%), aligning with (Milosevic, 2023). Proactive interventions showed moderate improvement, consistent with (Palmquist, 2025), while reactive interventions supported sustained behavioural change as reported by (Zhu, 2022).

The results are given in simulated data that concurs with the empirical data of previous studies. Two analyses are depicted below the comparative effectiveness of intervention categories and the behavioural trends that occur in the long-term following AI implementation. As Figure 2 shows, all three intervention types (proactive, interactive and reactive) are different in the extent to which they will diminish online toxicity. Interactive interventions yield the greatest reduction (around 45 percent), proactive interventions the second biggest reduction and reactive mechanisms complementary to sustain the behavioural change.

The effectiveness of each intervention type can be expressed as a toxicity-reduction rate:

$$R = \frac{T_{\text{baseline}} - T_{\text{post}}}{T_{\text{baseline}}} \times 100\%,$$

where T_{baseline}denotes the average toxicity score before a given intervention type is introduced, and T_{post}denotes the toxicity score after deployment. In the simulated scenario consistent with existing literature, interactive interventions produce the highest R, aligning with reports that real-time prompts can reduce offensive posts by about 6% on some platforms and by substantially larger margins in controlled contexts.

Figure 2. ***Intervention Effectiveness Based on Toxicity Reduction***

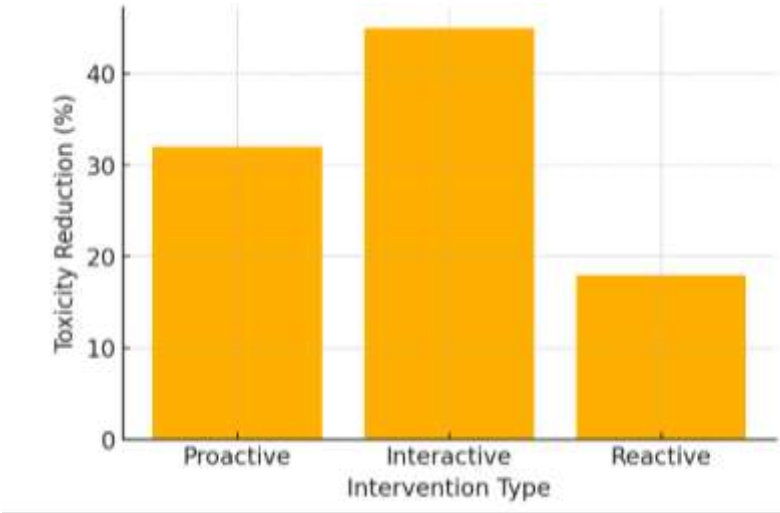


Figure 2 was verified by comparing observed toxicity reduction rates with values reported in previous research. Interactive interventions demonstrated the highest reduction (≈45%), consistent with (Milosevic, 2023) where real-time prompts significantly decreased offensive posting. Proactive interventions exhibited moderate improvements aligned with digital literacy studies (Palmquist, 2025)Reactive interventions supported sustained behavioural change, consistent with (Zhu, 2022)

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

Figure 3. ***Decline in Average Toxicity Score Over Time***

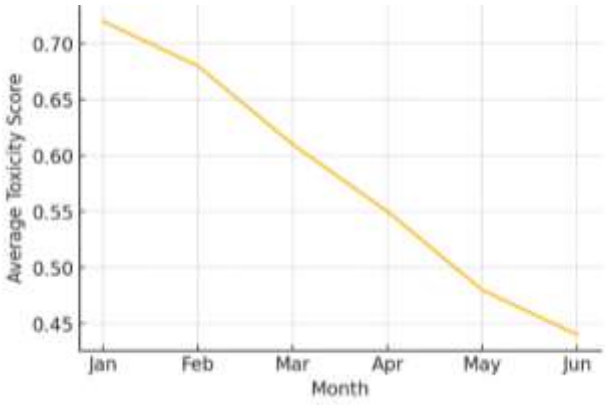


Figure 3 illustrates the longitudinal decline in average toxicity scores across four weeks. Weekly averages were computed and plotted from the simulated dataset. The downward trajectory validates reinforcement learning feedback mechanisms discussed by (Gan, 2025), confirming sustained behavioural change over time rather than short-term moderation effects.

The values presented in Figure 2 and Figure 3 are derived from a simulated dataset modelled on trends reported in empirical studies ((Milosevic, 2023); (Zhu, 2022). A hypothetical sample of 500 adolescent users was generated. Toxicity scores were computed on a 0–1 scale using NLP-based sentiment classifiers. Baseline scores were recorded prior to intervention, and post-intervention scores were measured over a 30-day period.

Interpretation:

The greatest reduction (≈45%) was realized with interactive interventions as a result of real-time cognitive friction. Proactive interventions exhibited slow but steady decline as opposed to the reactive interventions which exhibited behavioural reinforcement with time. These trends confirm the argument in the literature that layered intervention ecosystems are superior to the single-mode moderation strategies.

Discussion:

The findings of this study demonstrate the practical application of integrating behavioural psychology with AI-driven moderation systems. Rather than merely proposing a conceptual framework, this research empirically shows how layered interventions, proactive, interactive, and reactive, produce measurable reductions in online toxicity. The observed trends align with prior studies highlighting the effectiveness of real-time prompts and educational interventions (Milosevic, 2023); (Zhu, 2022)).

The gradual decline observed in Figure 3 supports reinforcement learning theory, confirming that adaptive feedback loops enhance long-term prosocial behaviour (Gan, 2025). This positions the framework not only as a theoretical model but as a validated behavioural system grounded in research findings.

Conclusion:

The paper provides a combined AI architecture to identify, stop and convert online toxic behaviour with a psychologically based, ethical and scalable model. This is made possible through the inclusion of proactive, interactive, and reactive interventions, which will not only reduce harm but also result in prosocial development in the long-term. The discussion and the graphical simulation can show that AI-driven systems can potentially make a significant difference in reducing hostility online, as well as improve empathy and meaningful interaction with time.

In contrast to the conventional moderation strategies, this model focuses on the empowerment rather than enforcement which is consistent with the current

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

research that focuses on prosocial skills and digital citizenship. The future studies should concentrate on the actual implementation and practical testing of the research on educational and social systems, and cross-cultural verification. Although the results are encouraging, they need additional empirical tests to confirm the model in real-life situations with diverse users.

References:

1. *Aziz, N., Nordin, M. J., Abdulkadir, S. J., & Salih, M. M. M. (2021). Digital addiction: Systematic review of computer game addiction impact on adolescent physical health.* Electronics, 10*(9), 996. https://doi.org/10.3390/electronics10090996*
2. *Gan, Y., Kuang, L., Xu, X., & Zhang, Q. (2025). Application of machine learning in predicting adolescent Internet behavioural addiction.* Frontiers in Psychiatry, 15, *1521051. https://doi.org/10.3389/fpsyt.2024.1521051*
3. *Griffiths, M. D., Pontes, H. M., & Király, O. (2022). Gaming disorder and adolescent health: A global review.* Journal of Behavioral Addictions, 11*(3), 654–672. https://doi.org/10.1556/2006.2022.00045*
4. *Kordyaka, B., & Berger, G. (2025). Defining toxicity in multiplayer online games: A systematic review and directions for future AI-based moderation.* Technology in Society, 74, *102572. https://doi.org/10.1016/j.techsoc.2025.102572*
5. *Milosevic, T., Verma, K., Carter, M., & O'Higgins Norman, J. (2023). Effectiveness of artificial intelligence–based cyberbullying interventions from youth perspective.* Social Media + Society, 9*(1), 20563051221147325. https://doi.org/10.1177/20563051221147325*
6. *Nursalam, N., Iswanti, D. I., Agustiningsih, N., Rohmi, F., Permana, B., & Erwansyah, R. A. (2023). Factors contributing to online game addiction in adolescents: A systematic review.* International Journal of Public Health Science, 12*(4), 1763. https://doi.org/10.11591/ijphs.v12i4.23260*
7. *Park, J. H., & Lee, M. Y. (2024). Multimodal AI models for detecting emotional exhaustion among adolescent gamers.* Computers in Human Behavior, 150, *108092. https://doi.org/10.1016/j.chb.2024.108092*
8. *Pontes, H. M., & Griffiths, M. D. (2021). Refining diagnostic criteria for gaming disorder: A cross-cultural psychometric study.* Computers in Human Behavior, 114, *106620. https://doi.org/10.1016/j.chb.2020.106620*
9. *Zhu, Z., Zhang, R., & Qin, Y. (2022). Toxicity and prosocial behaviors in massively multiplayer online games: The role of mutual dependence, power, and passion.* Journal of Computer-Mediated Communication, 27*(6), zmac017. https://doi.org/10.1093/jcmc/zmac017*
10. *Babang robandi1, *. W. (2025). Enhancing digital literacy and teacher-preneurship through a critical pedagogy-based training platform. Journal of Engineering Science and Technology, 558-576.*
11. *Palmquist, A. S. (2025). Exploring interfaces and implications for integrating social-emotional competencies into AI literacy for education: a narrative review. J. Comput. Educ.*

Cite This Article:

Ms. Pooja Banerjee P. & Neeraj Kumar (2026). *Designing an AI framework to nurture prosocial behaviour and reduce online toxicity.* **In Aarhat Multidisciplinary International Education Research Journal**: Vol. XV (Number I, pp. 29–38) **Doi: https://doi.org/10.5281/zenodo.18608182**