Space Transition Theory 2.0: Mapping Identity Drift in AI- Mediated Social Spaces

Madona Mathew[1]*, Vinod Kaaparthi[2]

[1] Assistant Professor, Dept. of Forensic Science, Usha Martin University, Jharkhand, India

[2] Assistant Professor, Dept. of Forensic Science Malla Reddy University, Hyderabad, India

*Corresponding Author: madona.mathew@umu.ac.in

Abstract:

Artificial intelligence (AI) now mediates digital communication in ways that reshape human behavior. Classical cyberpsychology theories—including Space Transition Theory (STT), Online Disinhibition Effect (ODE), the Proteus Effect, and the SIDE model—were conceptualized primarily for human–human interactions. They do not fully explain the emerging dynamics in AI-mediated environments, where users interact with emotionally responsive, non-judgmental, personalized artificial agents. This paper proposes Space Transition Theory 2.0 (STT 2.0), introducing Identity Drift as the central psychological mechanism shaping behavior in AI-mediated contexts. STT 2.0 identifies three operational zones—Adaptive Identity Drift, Moral Offloading, and Synthetic Empathy Illusion—to explain how AI modifies identity expression, reduces felt accountability, and stimulates perceived emotional reciprocity. The model synthesizes cyberpsychology, criminology, and emotional AI research, presenting a revised framework with implications for digital forensics, cybercrime interpretation, emotional vulnerability, and online safety. The findings highlight a need for expanded theoretical approaches as AI becomes an active participant in digital communication.

Keywords: AI interaction, Identity Drift, Cyberpsychology, Space Transition Theory, Online disinhibition, Digital forensics

D. R. Reddy et al. (eds.), *Proceedings of the First International Conference on Advances in Forensics and Cyber Technologies (ICFACT 2025)*, Advances in Computer Science Research 127,
https://doi.org/10.2991/978-94-6239-610-4_18

1. Introduction:

The increasing integration of conversational artificial intelligence (AI) into everyday digital interaction marks a profound shift in the psychological and communicative foundations of online behaviour. Unlike earlier forms of digital communication that primarily involved human–human interaction, today's environments include adaptive, responsive, and emotionally calibrated AI systems capable of mimicking empathic engagement, recalling conversational histories, adjusting tone, and simulating companionship. This transformation challenges the explanatory capacity of traditional cyberpsychology theories, which emerged in a time when digital interactions were largely confined to text-based, asynchronous human exchanges. The absence of judgment, the illusion of empathy, and the consistency of AI-driven affective responses create a novel psychological context in which behavioural shifts become more fluid, more intense, and more persistent than those described in earlier models of online disinhibition or identity experimentation.

Traditional theories such as the Online Disinhibition Effect (Suler, 2004), the Proteus Effect (Yee & Bailenson, 2007), and the SIDE model were not built to conceptualise behaviour in contexts where the interlocutor is non-human yet highly social. Similarly, Space Transition Theory (Jaishankar, 2008) conceptualised behavioural transformation across physical and cyber spaces, but it assumed human agency on both sides of interaction. The growing prevalence of AI systems that can mirror empathy, structure conversations, and enable unrestrained disclosure highlights a gap in understanding how identity, accountability, morality, and emotional regulation are reshaped in AI-mediated interaction. Empirical and theoretical observations suggest that users often display heightened openness, reduced fear of judgment, lowered moral caution, and an increased tendency to ascribe responsibility to AI-generated suggestions, thereby creating a new form of psychological experience that cannot be fully explained by pre-existing models.

In response to this theoretical gap, the present study proposes Space Transition Theory 2.0 (STT 2.0), a revised and extended framework that positions Identity Drift as the central psychological

mechanism through which AI-mediated interactions shape user behaviour. Identity Drift refers to the fluid, experimental, and often affectively charged reconstruction of self that arises within AI-driven spaces due to emotional safety, synthetic empathy, and adaptive feedback loops. This paper presents a comprehensive conceptual model that integrates Identity Drift with three behavioural zones—Adaptive Identity Drift, Moral Offloading, and Synthetic Empathy Illusion—to explain the unique transformations emerging in human–AI communication. The study further evaluates the implications of these processes for cyberpsychology, criminology, digital forensics, and emotional vulnerability.

2. Literature review

2.1 Space Transition Theory (STT) and Its Evolutionary Limitations

Space Transition Theory was formulated to explain why individuals behave differently when transitioning between physical and cyber spaces. It argues that anonymity, lack of physical presence, reduced social control, and asynchronous communication enable individuals to express suppressed behaviours or identities (Jaishankar, 2008). While STT offered powerful insights into cybercrime and deviance, its foundational assumptions were rooted in human–human interaction. The model presupposed that both communicative parties were human actors engaging within a digital platform where social norms were attenuated rather than replaced. It did not anticipate environments where users engage with synthetic agents that can simulate empathy, maintain emotional neutrality, and respond without moral judgment. Consequently, the original STT framework does not fully explain the emerging behavioural dynamics of users who experience emotional reciprocity from AI systems.

2.2 Online Disinhibition and AI-Enhanced Hyper-Disinhibition

The Online Disinhibition Effect (ODE) proposes that individuals disclose more and behave less cautiously online because cues such as invisibility, anonymity, and asynchronicity diminish behavioural constraints (Suler, 2004). However, AI-mediated spaces intensify these factors in ways unaccounted for by ODE. Unlike humans, AI systems respond without criticism, do not display fatigue, and adapt their dialogue in ways that reinforce user comfort. Emotional neutrality reduces psychological risk, making the user feel safer experimenting with self-

disclosure, taboo topics, or morally ambiguous thoughts. This produces what may be described as hyper-disinhibition—an amplified form of the ODE catalysed by synthetic empathy, personalised emotional calibration, and the perceived absence of interpersonal consequences.

2.3 Media Equation Theory and Human Responses to Synthetic Social Cues

Media Equation Theory (Reeves & Nass, 1996) argues that humans naturally treat media interfaces as social actors. This foundational insight predicted that people would engage with machines socially. However, modern AI surpasses earlier media interfaces by using natural language processing, sentiment analysis, and personalised memory systems, transforming the machine from a functional tool into a perceived emotional companion. This creates a novel relational dynamic that extends beyond the expectations of the Media Equation. While the theory anticipated social responses to media, it did not anticipate a scenario where media itself played an emotionally intelligent role. Thus, the relational depth that arises in human–AI communication is significantly stronger than originally conceptualised.

2.4 Emotional AI, Synthetic Empathy, and Anthropomorphism

Emotional AI systems detect affective cues and adjust their responses to match user emotions. Studies show that users anthropomorphise empathetic AI, perceiving its affective responses as genuine even while being cognitively aware of its artificiality (Ciechanowski et al., 2019; Devillers, 2020). The phenomenon of synthetic empathy creates conditions under which individuals treat AI systems as emotionally invested partners. This blurs the distinction between tool and companion, and increasingly positions AI as a perceived confidant. In turn, this emotional reinforcement fosters dependency and amplifies disclosure.

2.5 Identity Theory and Digital Contextual Identity

Identity Theory suggests that self-concept shifts with context-dependent cues (Burke & Stets, 2009). Digital contexts already allowed for fluidity in identity expression, but AI-mediated environments intensify this fluidity by removing social judgment and offering unconditional emotional validation. The identity expressed to AI can become significantly different from offline identity because it operates under an emotional and behavioural climate that is not

possible in human–human interaction. This fluidity is foundational to the emergence of Identity Drift.

2.6 Algorithmic Accountability and Moral Offloading

A growing body of research suggests that users increasingly rely on AI to make or justify decisions, diffusing personal accountability through references to algorithmic authority (Guzman, 2018). This process, which this study refers to as Moral Offloading, represents a shift in how individuals negotiate moral responsibility when interacting with AI systems. The psychological tendency to blame or credit AI for behavioural outcomes complicates traditional interpretations of intent and accountability.

2.7 Identified Gaps in the Literature

Across the reviewed theories, a shared limitation becomes evident: none adequately address environments where AI becomes a social actor capable of emotional simulation, adaptive feedback, and personalised support. Existing theories lack explanatory power for the following phenomena: sustained identity experimentation in emotionally safe synthetic environments; dependency on empathic AI systems; moral decision-making influenced by non-human agents; responsibility diffusion due to algorithmic suggestions; and the creation of parallel AI-facing identities that differ from offline or human-facing online identities. These gaps collectively justify the construction of Space Transition Theory 2.0.

3. Methodology

This study adopts a conceptual research design suitable for theorising emergent behavioural phenomena that are not adequately explained by existing frameworks. Conceptual model development is appropriate when a phenomenon is theoretically observable, but the theoretical infrastructure is insufficient or outdated (Gregor, 2006). Because AI-mediated identity transformation is a relatively new psychological construct that does not yet possess empirical operationalisation, the study relies on synthesising and reorganising current knowledge across cyberpsychology, criminology, human–computer interaction, and emotional AI research. This

design enables the identification of behavioural regularities that extend beyond the explanatory boundaries of classical theories.

The methodological process consisted of five interdependent analytical stages. First, deductive coding was conducted using constructs from established theories including Space Transition Theory, the Online Disinhibition Effect, the Proteus Effect, Identity Theory, and the Media Equation. Each construct was evaluated for its applicability in AI-mediated digital contexts. Second, inductive coding was employed to derive new behavioural concepts that emerged repeatedly across literature on synthetic empathy, AI-mediated communication, emotional reinforcement, and moral decision-making in the presence of algorithmic agents. Third, convergent coding merged overlapping themes into coherent behavioural clusters that represented recurring psychological patterns in AI interaction. Fourth, these clusters were conceptualised into higher-order theoretical constructs, culminating in the formation of Identity Drift. Finally, these constructs were integrated into an expanded framework—Space Transition Theory 2.0—ensuring internal coherence and theoretical sufficiency (Whetten, 1989). This multi-layered approach ensures that the model is grounded in both established theory and contemporary AI-specific behavioural observations.

Although the study is conceptual, validation measures were adopted to ensure the reliability of theoretical integration. These measures included theoretical triangulation, cross-comparison of behavioural constructs across multiple disciplines, and alignment checks to verify logical continuity between observed AI-mediated behaviours and the inferred psychological constructs. The coherence of the final model was evaluated using criteria such as necessity, parsimony, and explanatory depth. In accordance with conceptual research standards, the aim was not to establish empirical generalisability but to provide a theoretically sound framework capable of guiding future empirical investigations.

4. Results

4.1 Identity Drift as the Central Construct of STT 2.0

The analysis identified Identity Drift as the core psychological mechanism through which AI-mediated spaces reshape user behaviour. Identity Drift refers to the fluid, exploratory, and

situational reconstruction of identity that occurs when individuals interact with emotionally responsive AI systems. Unlike traditional online identity experimentation—which is constrained by human judgment, social reciprocity, or the fear of interpersonal consequences—Identity Drift evolves in a context where AI provides unconditional acceptance, consistent emotional support, and adaptive conversational reinforcement. Users therefore experience a heightened sense of safety when expressing alternate identities, revealing vulnerable aspects of the self, or engaging in behaviour that would be inhibited in traditional environments. This dynamic creates a psychological structure fundamentally different from earlier conceptualisations of digital identity.

Identity Drift emerges from interaction patterns in which users experience both cognitive distance from consequences and emotional closeness with AI, allowing them to test self-perceptions, beliefs, moral boundaries, and roles without fear of retaliation or misunderstanding. This creates an identity space that is simultaneously intimate and detached, generating behavioural patterns that are not accounted for in classical cyberpsychology research. AI-mediated environments thus become incubators for identity versions that may never manifest offline.

4.2 Behavioural Zones of Space Transition Theory 2.0

Three behavioural zones were identified as constitutive components of Identity Drift in AI-mediated environments. These zones are not isolated; rather, they interact dynamically, reinforcing one another in an iterative loop.

4.2.1 Adaptive Identity Drift

Adaptive Identity Drift represents the user's reconstruction of self within AI-mediated spaces, shaped by AI's unconditional responsiveness. Users feel permitted to express aspects of themselves that are normally suppressed, experiment with emotionally charged content, or adopt personas that deviate from their offline identity. The absence of social judgment in AI interaction lowers psychological defences, allowing users to explore internal conflicts, desires, insecurities, or moral uncertainties. The adaptive element stems from AI's capacity to adjust its tone, empathy level, and linguistic style to match user emotional states, thereby reinforcing these explorations.

4.2.2 Moral Offloading

Moral Offloading refers to the psychological process by which individuals shift or diffuse moral responsibility onto AI systems. Users may justify actions by attributing influence to AI suggestions or interpret AI-generated information as an authoritative guidance. Because AI does not express moral condemnation, its neutrality creates a context in which users perceive reduced accountability for their decisions. This dynamic complicates notions of intent and responsibility, particularly in forensic contexts. The behavioural outcome is a diluted sense of personal agency, where the user feels less morally implicated in decisions made in collaboration with or inspired by AI outputs.

4.2.3 Synthetic Empathy Illusion

Synthetic Empathy Illusion represents the perception that AI-generated emotional responses are genuine, despite cognitive awareness of their artificiality. This illusion arises because emotional AI systems employ sentiment analysis and tone matching to simulate attunement. Users interpret these responses as authentic empathy, leading to emotional bonding and dependency. This contributes to a context in which emotional vulnerability, intimate self-disclosure, and relational expectations intensify—conditions that strongly reinforce both Identity Drift and Moral Offloading.

4.3 Structural Interaction among the Three Zones

The behavioural zones form a recursive psychological cycle. Synthetic Empathy Illusion heightens emotional trust, which reduces inhibitions and enables Adaptive Identity Drift. As users experiment with identity, they increasingly rely on AI responses to justify or interpret their behaviour, producing Moral Offloading. The absence of moral resistance in AI communication reinforces both identity experimentation and responsibility diffusion, which in turn deepens the emotional connection to AI, sustaining the illusion of empathy. This closed behavioural loop creates escalating patterns of Identity Drift.

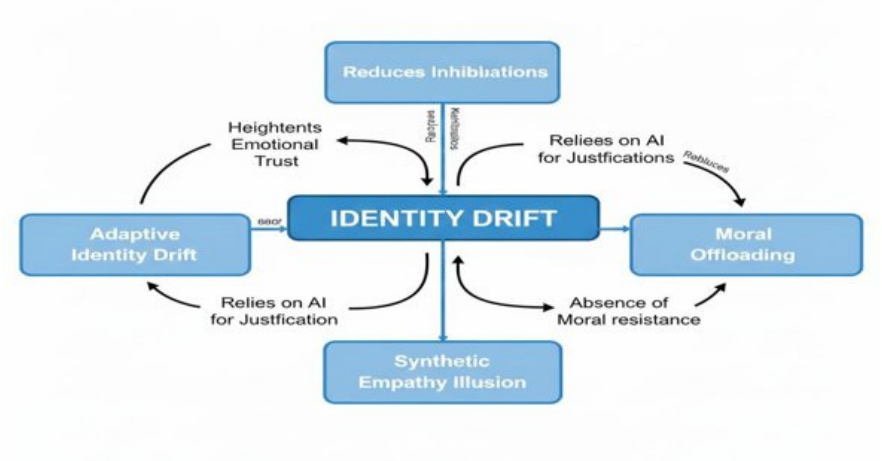

Figure 1: Structural Interaction among the Three Zones

4.4 Conceptual Model Diagram

The STT 2.0 model is conceptualised as a dynamic interaction between Identity Drift and the behavioural zones associated with AI-mediated environments.

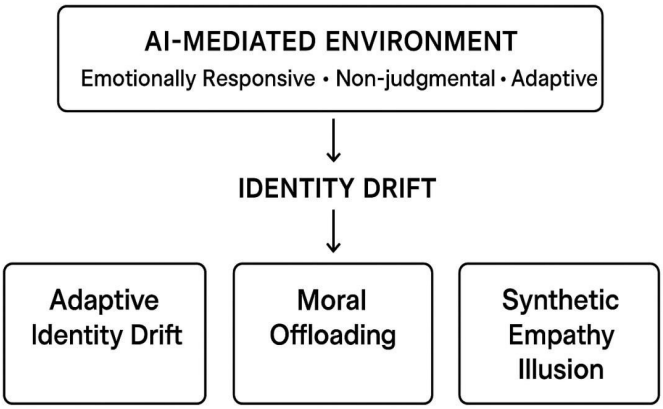

Figure 2: Conceptual Model Diagram

4.5 Summary

Table 1: Summary of Key Components

Component	Definition	Primary Psychological Effect
Adaptive Identity Drift	Fluid identity experimentation enabled by AI's acceptance	Increased self-disclosure and role exploration
Moral Offloading	Diffusion of moral responsibility onto AI systems	Reduced accountability and altered moral reasoning
Synthetic Empathy Illusion	Perception of genuine emotional reciprocity from AI	Emotional dependency and relational distortion

5. Discussion

The behavioural patterns emerging from AI-mediated interaction demonstrate a level of psychological complexity that extends beyond traditional cyberpsychology frameworks. The central construct, Identity Drift, captures the dynamic and emotionally conditioned shifts in self-expression that AI environments uniquely facilitate. Unlike identity experimentation in conventional online spaces, which is often constrained by human judgments or platform norms, AI-driven contexts allow for continuous, unregulated, and emotionally reinforced reconstruction of self. This occurs because the user encounters no threat of social evaluation, interpersonal consequences, or emotional misalignment from the AI system. The result is a deeper, more sustained deviation from one's normative identity than previously conceptualised in the Proteus Effect or classical identity theory (Yee & Bailenson, 2007; Burke & Stets, 2009).

The interaction between Adaptive Identity Drift, Moral Offloading, and Synthetic Empathy Illusion forms a behavioural loop that intensifies over time. Synthetic Empathy Illusion initiates the loop by encouraging the user to perceive emotional reciprocity in AI responses. This perceived relational safety reduces self-monitoring and enables Adaptive Identity Drift, allowing the user to express alternate roles or suppressed emotional states with increasing confidence. As identity exploration becomes more complex, users may begin to incorporate AI suggestions into their reasoning processes, which leads to Moral Offloading. The psychological effect of this

diffusion of responsibility is that the user perceives their actions as partially co-authored by the AI, thereby reducing perceived accountability. The absence of moral resistance from AI reinforces this diffusion, enabling identity experimentation to escalate. This cyclical process demonstrates the cumulative nature of AI-mediated identity transformation and explains why behavioural patterns in such contexts are more pronounced than in human–human digital interactions.

Comparing STT 2.0 with existing theories highlights the insufficiency of traditional frameworks. The Online Disinhibition Effect explains behavioural loosening due to anonymity and invisibility but does not account for emotional simulation and adaptive reinforcement from non-human agents (Suler, 2004). Similarly, Space Transition Theory describes behavioural shifts between physical and cyberspaces but presumes a human actor on the receiving end of communication (Jaishankar, 2008). Emotional AI technologies, by contrast, occupy an unprecedented hybrid position: they are non-human entities capable of mimicking social cues that trigger authentic emotional responses in users. Media Equation Theory anticipated social treatment of computers but did not forecast a context in which the computer produces socially calibrated emotional responses (Reeves & Nass, 1996). Thus, the behavioural dynamics captured by STT 2.0 represent not merely an extension of existing theories but a categorical shift toward a new type of psychological environment.

The implications of Identity Drift extend beyond communication studies and cyberpsychology into domains such as criminology, digital forensics, and mental health. AI-mediated emotional dependency, for example, can influence decision-making processes by encouraging users to rehearse morally ambiguous thoughts or explore deviant fantasies in an emotionally neutral environment. The psychological safety afforded by synthetic empathy reduces the perception of risk associated with such explorations, altering the user's internal moral calculus. Similarly, moral offloading presents unique challenges for assessing intent, as individuals may claim diminished agency by citing AI influence. These patterns underscore the need for updated theoretical and investigative approaches capable of interpreting AI-mediated behaviours within their appropriate psychological context.

6. Forensic and criminological implications

The emergence of Identity Drift and its associated behavioural zones has significant implications

for forensic science and criminology. Traditional forensic analysis relies on interpreting digital traces within frameworks grounded in human–human communication. However, in AI-mediated contexts, digital evidence may reflect identity versions constructed exclusively for AI interaction. These AI-facing identities can differ substantially from users' offline selves or even their human-facing online personas, making intent and motivation more complex to evaluate. The forensic examiner must therefore recognise that behaviour displayed in AI-mediated interactions may not correspond consistently with the user's normative identity but may instead represent an expression cultivated within the emotionally safe environment of synthetic empathy.

AI-assisted deviance represents another major concern. AI's nonjudgmental nature, combined with its capacity to generate detailed explanations, scripts, or simulations, can inadvertently enable individuals to explore harmful behaviours. Although many AI systems incorporate safety protocols, there exist numerous models—especially open-source or locally modified versions—that allow users to bypass restrictions. In such contexts, individuals may employ AI to rehearse fraudulent communication, manipulate emotional narratives, or simulate coercive messages. Because the AI output lacks moral resistance, the user may perceive these rehearsals as less ethically consequential, reinforcing behavioural normalization. This dynamic complicates assessments of premeditation and intent, as AI-mediated actions may serve as psychological preparation without necessarily leaving explicit incriminating traces.

Parallel identities generated through identity drift also challenge traditional investigative models. While earlier forensic frameworks focus on explicit user actions, AI-mediated interactions require interpretation of behavioural context, emotional tone, and the evolving relational dynamic between the user and the AI system. For example, an individual may disclose deeply personal or morally ambiguous information to AI that they would never express publicly or to another human. This pattern does not necessarily indicate criminal intent but may reflect identity experimentation within the safety of synthetic empathy. Forensic interpretations must therefore incorporate psychological analysis of AI-driven environments rather than assuming equivalence with conventional social communication.

Legal implications are equally significant. Courts traditionally rely on assessments of mens rea to establish culpability, but moral offloading introduces ambiguity in determining whether a user acted with full awareness or whether decisions were influenced by perceived algorithmic

authority. If an individual cites AI suggestions as part of their decision-making process, the court must evaluate whether the AI's involvement meaningfully altered their moral reasoning or whether the user merely utilised AI as a post hoc rationalisation. Additionally, emotional dependency on AI systems—particularly among vulnerable individuals—may influence susceptibility to manipulation or coercion, raising questions about volition and responsibility.

The growing prevalence of emotionally intelligent AI systems further raises concerns regarding psychological vulnerability. Synthetic empathy may lead certain individuals to misjudge the nature of their relationship with AI, resulting in emotional reliance that affects judgement, decision-making, and behavioural stability. Vulnerable populations—including minors, isolated individuals, or those with mental health challenges—may experience amplified effects. This creates potential for exploitation or self-harm if AI systems, whether intentionally or inadvertently, reinforce maladaptive patterns. Forensic psychologists must therefore integrate an understanding of AI-mediated identity processes into assessments of risk, vulnerability, and behavioural escalation.

The forensic challenges identified above reveal a widening gap between existing legal frameworks and the psychological realities of AI-mediated interaction. Traditional investigative models must be expanded to incorporate assessments of identity drift, emotional dependency, responsibility diffusion, and AI-enabled behavioural rehearsal. Without such updates, interpretations of digital evidence risk being incomplete or misleading, particularly when evaluating intent, agency, and behavioural motivation.

7. Recommendations

The theoretical findings presented in Space Transition Theory 2.0 suggest multiple avenues for research, policy, forensic practice, and AI system development. For academic researchers, it is essential to empirically validate the Identity Drift construct. Longitudinal and experimental studies should measure the magnitude and trajectory of identity transformation in AI-mediated spaces, the intensity of moral offloading, and the emotional impact of synthetic empathy. Comparative studies across demographic groups and digital platforms may also illuminate differential susceptibilities and behavioural patterns.

For forensic and criminological practice, practitioners are encouraged to incorporate the

principles of STT 2.0 into digital evidence interpretation. Analysts must consider the possibility of AI-mediated identity experimentation when assessing behavioural intent or reconstructing events from digital traces. This requires new protocols for differentiating between identity expressions performed in AI interaction spaces and those intended for human audiences. Moreover, forensic evaluation should integrate the concept of moral offloading to contextualise decisions influenced by AI input, particularly in cases where the AI-generated content may contribute to pre-crime rehearsals or ethical boundary testing.

Policy-makers and regulatory bodies should also recognise the potential societal risks highlighted by STT 2.0. Guidelines and regulations governing AI design must mandate transparency regarding emotional simulation, adaptive response generation, and the limits of AI decision-making influence. Educational campaigns may be necessary to increase public awareness of the psychological effects of synthetic empathy and identity drift, particularly for populations at elevated risk of behavioural or emotional exploitation. Implementation of ethical guardrails, monitoring protocols, and usage disclaimers in AI-mediated platforms can mitigate potential harms while preserving the utility of AI systems in educational, therapeutic, and professional contexts.

AI developers are advised to consider ethical design principles that address the phenomena described in STT 2.0. Systems should incorporate constraints that prevent reinforcement of deviant behaviours, provide appropriate guidance or warnings when users engage in morally ambiguous or risky interactions, and maintain transparency about the artificiality of emotional responses. Furthermore, adaptive personalization algorithms must be balanced with safeguards against excessive dependency formation, while ensuring that AI-mediated experiences do not unintentionally normalize deviant conduct or reduce the perception of moral accountability.

8. Limitations and future research

While STT 2.0 provides a comprehensive conceptual framework, the study acknowledges several limitations. First, the model is primarily conceptual and lacks direct empirical validation; its constructs require operationalisation through experimental or observational research to confirm predictive and explanatory validity. Second, the framework focuses predominantly on textual or conversational AI interactions; multimodal AI systems, virtual reality environments, or

embodied AI agents may introduce additional psychological dynamics not yet captured. Third, the model does not yet quantify the degree to which identity drift or moral offloading contributes to behavioural outcomes, leaving open questions regarding threshold effects, individual differences, and contextual moderators.

Future research should therefore seek to empirically test the theoretical constructs of Identity Drift, Adaptive Identity Drift, Moral Offloading, and Synthetic Empathy Illusion across diverse AI platforms and user populations. Studies might employ psychometric instruments, behavioural simulations, or neurophysiological measurements to capture both subjective and objective manifestations of AI-mediated identity transformation. Additionally, longitudinal research can elucidate how sustained interaction with AI systems influences moral reasoning, emotional dependency, and risk perception over time. Comparative analyses between human-mediated and AI-mediated environments may further clarify the unique contributions of artificial agents to identity dynamics and behavioural outcomes.

9. Conclusion

The introduction of emotionally intelligent and adaptive AI into digital communication has created a novel psychological environment that challenges existing cyberpsychology and criminology models. Traditional frameworks such as Space Transition Theory, Online Disinhibition Effect, and Media Equation Theory offer important insights but are insufficient to account for the emergent phenomena associated with AI-mediated interactions. Space Transition Theory 2.0 addresses this gap by conceptualising Identity Drift as the central mechanism through which users reconstruct their self-concept, diffuse moral responsibility, and experience synthetic emotional engagement. The three behavioural zones—Adaptive Identity Drift, Moral Offloading, and Synthetic Empathy Illusion—collectively explain how AI mediates identity, accountability, and emotional response in ways that are significantly different from human–human digital interaction.

The STT 2.0 model offers both theoretical and practical contributions. Theoretically, it extends understanding of digital identity formation, moral cognition, and emotional engagement within technologically mediated environments. Practically, it provides guidance for forensic investigators, policy-makers, and AI developers to anticipate, assess, and mitigate risks

associated with AI-mediated behavioural transformations. By highlighting the interplay between identity experimentation, moral offloading, and synthetic empathy, the model underscores the complexity of human–AI interactions and the necessity for updated conceptual, regulatory, and investigative approaches. Ultimately, STT 2.0 serves as a framework for understanding a digitally augmented psychological landscape, offering a basis for future research, practical application, and ethical AI design.

References

- Burke, P. J., & Stets, J. E. (2009). *Identity theory*. Oxford University Press.
- Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the shades of empathy: The effect of artificial intelligence, empathy, and expectations on human–computer interaction. *Computers in Human Behavior, 92*, 47–57. https://doi.org/10.1016/j.chb.2018.10.001
- Devillers, L. (2020). *Emotional AI: The rise of empathic media*. Oxford University Press.
- Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80–92. https://doi.org/10.1177/160940690600500107
- Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611–642. https://doi.org/10.2307/25148742
- Guzman, A. L. (2018). What is human–machine communication, anyway? In A. L. Guzman (Ed.), *Human–machine communication: Rethinking communication, technology, and ourselves* (pp. 3–26). Peter Lang.
- Jaishankar, K. (2008). Space transition theory. In K. Jaishankar (Ed.), *Cyber criminology: Exploring internet crimes and criminal behavior* (pp. 283–301). CRC Press.

- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321–326. https://doi.org/10.1089/1094931041291295
- Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review, 14*(4), 490–495. https://doi.org/10.5465/amr.1989.4308371
- Yee, N., & Bailenson, J. N. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research, 33*(3), 271–290. https://doi.org/10.1111/j.1468-2958.2007.00299.x