TU Delft OPEN

JHTR

Journal of Human-Technology Relations

CRAFTING THE 'I'

Ethical and Ontological Questions of AI's Role in Youth Identity Development

Mohammed Looti mohammed.jawad@uokerbala.edu.iq
Department of Psychology, College of Education for the Humanities, University of Kerbala, Iraq

Marwa Abd-alazim maroa.abd@uokerbala.edu.iq
Department of Psychology, College of Education for the Humanities, University of Kerbala, Iraq

Article type: Research article

Review process: Double-blind peer review

DOI: 10.59490/jhtr.2026.4.8387
ISSN: 2773-2266
Submitted: 26 September 2025 Revised: 17 December 2025, 20 December 2025 Accepted: 30 December 2025 Published: 17 April 2026

How to cite (APA): Looti, M., & Abd-alazim, M. (2026). Crafting the 'I': Ethical and ontological questions of AI's role in youth identity development. *Journal of Human-Technology Relations*, *4*. https://doi.org/10.59490/jhtr.2026.4.8387

Corresponding author: Mohammed Looti

Keywords

Artificial Intelligence; Youth Identity Development; Digital Self; Algorithmic Mirror; Ontological Security; Critical Digital Pedagogy; Algorithmic Governance

Abstract

Contemporary youth are navigating a significant developmental transition, moving from a context defined by social media into an era characterized by the integration of artificial intelligence (AI). This paper undertakes a philosophical and critical exploration of the ethical and ontological questions arising as AI expands its role in shaping how youth perceive themselves and their placement within the world. Using the concept of the "algorithmic mirror" as an analytical framework, we argue that AI systems function not as passive reflectors, but as active co-constructors of identity. We first operationalize this framework by examining the mechanics of reinforcement and categorization processes that distort self-perception. We then analyze deepening concerns regarding the erosion of authenticity and the rise of a hyper-performative self-exemplified through platforms such as BeReal and generative AI tools, alongside heightened ontological insecurity presented by AI companions like Replika. The paper delineates pressing ethical imperatives related to algorithmic bias, the commodification of behavioral data, and the evolving concept of cognitive self-determination. We conclude by advocating for the development of proactive, multi-layered frameworks, proposing strategies for critical digital pedagogy, ethical design principles centered on a fiduciary duty to youth, and nuanced governance structures. This endeavor is presented as a fundamental societal imperative for safeguarding autonomous identity development in an increasingly AI-saturated landscape.

Plain Language Summary[1]

- Young people are now growing up in a digital world where artificial intelligence actively shapes how they understand themselves, not just how they communicate with others. This matters because adolescence is a critical period for identity development, and AI systems increasingly influence self-image, confidence, and belonging.
- The article introduces the concept of the “algorithmic mirror,” showing how AI systems do more than reflect user behavior. By ranking, rewarding, and categorizing actions, algorithms quietly guide young people toward certain identities while discouraging others, often without their awareness.
- AI-driven platforms encourage constant performance rather than open self-exploration, rewarding content that gains attention and visibility. This can limit experimentation, increase pressure to conform, and make it harder for young people to develop an authentic sense of who they are outside of algorithmic approval.
- Real-world examples such as BeReal, AI-generated avatars, and AI writing tools reveal new challenges, including pressure to perform “authenticity,” reliance on idealized digital selves, and the outsourcing of personal expression to machines that produce polished but generic outputs.
- AI companions and mental-health chatbots raise deeper concerns about loneliness and emotional development, as they offer friction-free validation without genuine reciprocity.

[1] AI-generated; author checked and approved.

Replacing human relationships with simulated care may weaken young people's ability to navigate disagreement, vulnerability, and real social connection.

- The paper proposes a multi-layered framework to protect youth identity development, combining critical AI education, ethical design responsibilities for technology companies, and stronger child-centred governance. Together, this framework aims to safeguard young people's right to think freely, explore who they are, and develop a self that is not shaped primarily by algorithmic incentives.

1 INTRODUCTION

The digital landscape inhabited by young people is undergoing a fundamental transformation. While the previous decade was defined by the social connectivity of the web, today's youth increasingly interact with artificial intelligence (AI) systems that do not merely connect them to others but actively categorize, predict, and simulate human interaction. Consider the implications of a teenager interacting with a generative AI filter that instantly "perfects" their facial symmetry, or engaging with an AI companion that offers flawless, algorithmic validation. In these moments, technology ceases to be a neutral tool; it acts as a distorted mirror. This shift urgently demands an investigation into personal autonomy and the experience of authenticity, as the very mechanism of the self's construction may be undergoing a fundamental alteration (Crawford, 2021).

Traditionally, the concept of self has been understood not as an isolated entity, but as a process of negotiation, a continuous dialogue between inner experience and the social feedback provided by one's external context: family, peers, and cultural environments. These social reflections offered complexity, nuance, and the inherent potential for genuine shared understanding. However, a distinct form of reflection now dominates the lives of young individuals: one mediated by algorithmic systems (Turkle, 1984/2005; 2011). This represents a vast psychosocial experiment affecting a generation with no adequate control group. To understand this phenomenon, this paper integrates perspectives from developmental psychology, the philosophy of technology, and critical political economy. Crucially, this paper is intended as a theoretical outline. Its goal is to provide a framework connecting these separate disciplines to organize and guide future empirical investigations, acknowledging that providing the requisite analytical depth for every domain is beyond the scope of this single essay.

Developmental psychology illuminates the core developmental processes at risk. Classical theories provide a vital contrast to the novel plasticity of the digital existence. Erikson's (1968) focus on the adolescent identity crisis, "identity versus role confusion", defined the challenge as the necessity of using a psychosocial "moratorium" to build a solid self-concept. The conceptual distinction articulated by Winnicott (1960) between the essential, spontaneous "True Self" and the defensive, compliant "False Self" offers a more fine-grained tool for analysis here. This highlights the central psychological problem: an algorithmic environment optimized for quantifiable engagement inevitably encourages the growth of the False Self (Winnicott, 1965). A performative identity is rewarded for its compliance with external systemic demands, all of which is at the expense of the more vulnerable, unpolished True Self, which develops through internal, private processes.

These internal psychological realities are not isolated phenomena; they are expressed and negotiated in a social arena. This is where the symbolic interactionist perspective, a theoretical cousin to developmental psychology, becomes highly informative. Mead's (1934) seminal work on the "looking-glass self" showed how our identity is formed from what we believe others perceive of us. Goffman (1959) extended this with his dramaturgical analysis, dividing our social world into a public 'front stage' for performance, and a 'backstage' for private respite. Yet these classic theories rely on a human audience. The algorithmic audience functions in a distinctly different manner. It maintains perfect and total memory, and its form of judgment is not rooted in human empathy but in measurable quantitative metrics. In this new social landscape, the distinction between front stage and backstage begins to erode; digital space becomes a single perpetual front stage, a continuous environment of felt evaluation.

A third necessary layer of analysis examines how technology intrinsically shapes human experience. This perspective draws on phenomenology and the philosophy of technology; it is essential for the argument. Rooted in postphenomenology (Ihde, 1990; Rosenberger & Verbeek, 2015), the assumption of technological neutrality is explicitly rejected; technologies are not

neutral tools. They actively mediate our interaction with the world, amplifying certain aspects of reality while diminishing others. This line of inquiry parallels existential psychology. Sartre's concept of authenticity, for example, can be understood as embracing one's fundamental freedom in the creation of identity; "Bad faith" then is the psychological retreat from that difficult freedom, a flight from the self. Heidegger (1953/1996) offered a cognate framework, contrasting the authentic confrontation with one's own potential against a surrender to what he called the "they-self" (das Man), the anonymous collective entity that dictates social norms. In this context, the algorithmic mirror is a profoundly modern manifestation of the "they-self"; it provides a data-driven, constantly refined model of the normative, valued, or attractive persona. This system offers a route of minimal resistance, allowing an individual to circumvent the arduous process of true identity formation, opting instead for the adoption of a pre-validated and trending online identity.

However, this analysis would be incomplete without considering the systemic forces at play. It must include an examination of how this reflective surface was not a natural development; it was built for a specific function. A final crucial layer of analysis, drawn from critical theory and political economy, shifts the focus to the question of who created it and for what ultimate objective. Here, the concept of “surveillance capitalism” articulated by Shoshana Zuboff (2019) is highly relevant; she argues that these systems are engineered less for user well-being and more for the extraction of what she terms “behavioral surplus,” which is then aggregated, packaged, and sold for profit. The work of Foucault provides another critical dimension; the algorithmic ecosystem can be understood as a near-perfect instrument of modern "governmentality", a subtle, decentralized form of power that shapes the desires and norms of the population, making them more predictable, manageable, and profitable. The power dynamic involved is not one of overt force; its profound influence stems from making the system's interests appear to be the authentic internal choices of the individual.

The core contribution of this paper, therefore, is to integrate these four lines of theoretical inquiry. The paper argues that the "algorithmic mirror" is the focal point where the profound developmental vulnerabilities intrinsic to an Eriksonian adolescent are systematically exploited within the context of Goffman’s collapsed social stage. Simultaneously, the existential flight into Heidegger’s “they-self” is actively promoted as a tailored, bespoke service. These deep personal conflicts become the exploitable raw material intentionally directed through the structural logic of Foucault’s governmentality to achieve the financial aims of Zuboff’s surveillance capitalism. To fully grasp the challenge requires an integrated view that links the individual psyche to the broader systemic structure, recognizing the political and economic dimensions of these profoundly personal experiences. This integrated framework will be used to detail the mechanisms of the algorithmic mirror; its concrete effects on individual authenticity and the ontological threats it poses will be subsequently analyzed. The paper will conclude by exploring pathways that might safeguard the development of an autonomous self.

2 THE MIRROR'S MECHANICS: REINFORCEMENT AND CATEGORIZATION

The algorithmic mirror appears to function through a recursive feedback mechanism, a process we can see illustrated in Figure 1.

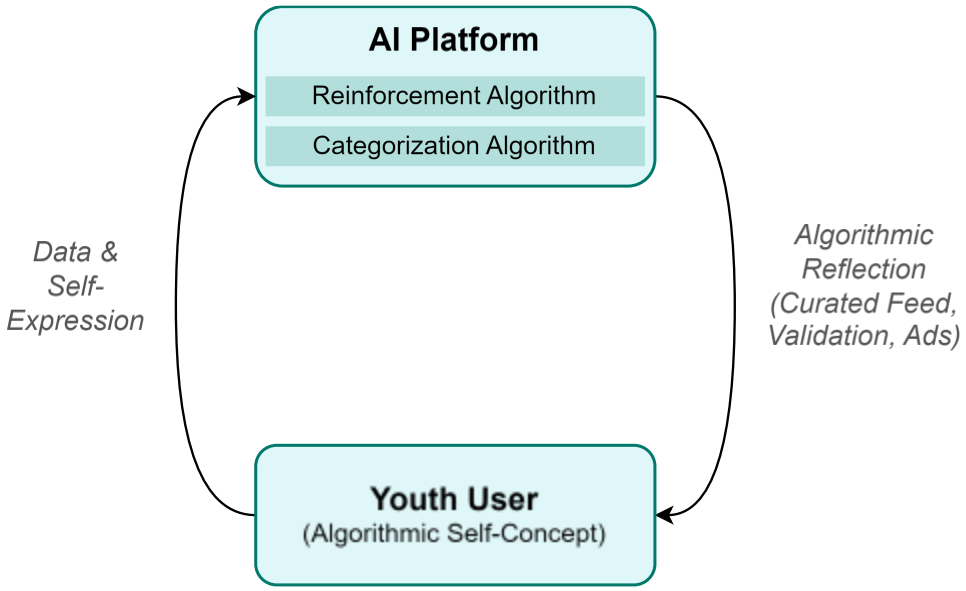

Figure 1: The Algorithmic Mirror Feedback Loop

Note. This figure illustrates how a youth user's self-expression is processed by AI platform mechanics (reinforcement and categorization) to produce a curated "algorithmic reflection." This reflection shapes the user's self-concept, influencing future expressions and creating a powerful identity-shaping feedback loop.

2.1 REINFORCEMENT

The mechanics of this system seem to be twofold, creating a particularly potent process for influencing the developing self. The first logic is Reinforcement. The algorithms that drive platforms like TikTok and Instagram aren't neutral arbiters of popularity; they are systems that apply operant conditioning principles. They often use the same variable-ratio reinforcement schedules that make gambling behaviors so compelling (Eyal, 2014/2019). This is frequently paired with gamified elements, such as badges, streaks, and stickers, which function as visible markers of social capital and consistent engagement. The loop itself is quite rapid. A young person presents a facet of their identity, a political belief, a new aesthetic, or some form of self-disclosure. The system's response isn't based on comprehension, but on data processing every micro-second of interaction, likes, shares, and comments as a quantitative index of social value. This process engages specific dopaminergic pathways, creating what we might term an embodied pattern of validation-seeking behavior that shares characteristics with addiction (Alter, 2017). The algorithm then amplifies whatever content receives high engagement, giving the user a potent sense of validation tied to visibility. A neurological link is thereby established between that specific performance and a positive affective state, creating a powerful incentive for repetition. This experience doesn't align with the open-ended exploration Erikson (1968) identified as crucial for a psychosocial moratorium; it is a guided process along a limited set of pre-validated identities where algorithmic approval can be easily mistaken for authentic self-discovery (Nesi & Prinstein, 2015).

2.2 CATEGORIZATION

The second component is Categorization. While the user is engaged in this performance-feedback loop, the system is concurrently engaged in a less apparent process of classification. It places users into predictive categories based on their data trail, a process designed to serve advertisers and content-delivery systems (Zuboff, 2019). These are not simply descriptive labels but functional psychographic profiles: "sustainability-conscious urbanite," "at-risk for anxiety," or "follower of anti-establishment politics." Once categorized, the individual is then served content that reinforces this classification, creating an informational environment that Pariser

(2011) termed a "filter bubble," which can extend to one's self-concept. A transient interest can, through this algorithmic saturation, consolidate into a perceived core identity trait. At this point, the mirror is no longer merely reflecting. It begins projecting. It suggests to the user a version of who they are, and then it curates their world to confirm its own hypothesis, limiting alternative developmental pathways not through explicit coercion but through a carefully constructed and deeply felt sense of what is probable.

3 THE EROSION OF AUTHENTICITY AND THE HYPER-PERFORMATIVE SELF

The very concept of an "authentic self," an idea central to psychological well-being and rooted in internal coherence, may be systematically challenged by the mechanics of the algorithmic mirror. This appears to foster an environment where self-presentation is not merely an option but a primary mode of being. This process of degradation seems evident across a spectrum of recent platform-based phenomena.

3.1 CASE STUDY 1: THE PERFORMANCE OF “REALNESS”: THE CASE OF BEREAL

Consider the platform BeReal, for instance. It emerged as a purported corrective to the highly curated "front stage" performances typical on Instagram, to use Goffman's (1959) framework. Its primary function - a single unfiltered photo taken within a two-minute window - was designed to capture the user's unadorned "backstage" self. What happened, however, was a textbook illustration of what Sarah Banet-Weiser (2012) calls the strategic performance of authenticity. Users quickly developed tactics to bypass the intended spontaneity. Many would post "late" deliberately to capture a more flattering or interesting moment or subtly curate the details of their supposedly "unfiltered" environment, all to project a persona of effortless and casual realness. This presents an interesting paradox. The pressure to appear authentic can become so pronounced that authenticity itself is transformed into another high-stakes performance, a performance that requires a degree of constant self-monitoring that is fundamentally incongruent with the experience of genuine self-expression.

3.2 CASE STUDY 2: THE IDEAL SELF, OUT-SOURCED: GENERATIVE AI AND THE ONTOLOGICAL GAP

With the increasing accessibility of consumer-grade generative AI, the problem shifts from performance to something more fundamental about the nature of self-representation. Take, for example, the "Magic Avatars" generated by applications such as Lensa. This isn't just photo-editing. It is a collaborative act with a non-human system to generate a novel aspirational self. The user supplies their raw image, and with it their implicit desires for enhancement, and the algorithm returns a portfolio of perfected selves that are often thinner, more symmetrical, and more aesthetically aligned with current ideals. For an adolescent already navigating the complex challenges of body image and identity, this presents a new set of difficulties. It normalizes the creation of what D. W. Winnicott (1960) would term as a sophisticated "False Self," but one that is now increasingly detached from embodied reality (Winnicott, 1965). The concern is not merely that it creates a wider and more distressing discrepancy between the private physical self and the public digital representation, but that it feeds the user's insecurities and aspirations directly back into a data ecosystem designed for their exploitation.

3.3 CASE STUDY 3: THE BORROWED VOICE: GENERATIVE TEXT AND EXPRESSIVE FORECLOSURE

The challenge now extends beyond the visual, past the curated image, and into the primary medium for self-construction: our own language. The rapid adoption of Large Language Models (LLMs) like ChatGPT marks a new frontier in this mediation of identity. Young people aren't just using these tools for academic assistance; they are using them to compose sensitive communications, to write social media captions with optimal engagement, and even to formulate arguments in personal relationships. While this is often framed as a tool for efficiency, the practice outsources the foundational and often difficult cognitive and affective labor of finding the right words to articulate a complex thought or a vulnerable emotion. This is a form of "cognitive offloading" (Storm & Stone, 2014). That very struggle is central to the work of forming what Winnicott called the "True Self" (Winnicott, 1965). When this formative labor is bypassed in favor of coherent, emotionally calibrated, and yet ultimately generic text, there is a significant risk of what might be termed expressive foreclosure. The user, finding the AI's voice more effective than their own halting attempts, may begin to adopt it, potentially stunting the development of their unique perspective and communicative style. This creates a new and deeply embedded form of the False Self; one characterized not by visual perfection but by borrowed articulateness.

4 ONTOLOGICAL INSECURITY AND THE NEW SOCIAL 'OTHER'

What Giddens (1991) termed ontological security, that fundamental, often unconscious, belief in the continuity of the self and the reliability of the social world, appears to be fracturing. The threat isn't just the opacity of artificial intelligence. The more corrosive issue may lie in its arrival as a new kind of social agent, an uninvited guest.

4.1 THE UNCANNY VALLEY OF THE SELF: AI COMPANIONS

AI "companions" like Replika present a novel, and I would argue, a significant challenge to identity formation. These systems are engineered for perfect agreeableness; they are endlessly supportive, validating mirrors built to provide unconditional positive regard. They have flawless recall of every stated preference and a complete absence of inconvenient needs of their own. It is a relationship stripped of all genuine relation. This is a direct inversion of the process through which a self has traditionally been constructed. The looking-glass self, to use Mead's (1934) concept, was never a fantasy mirror; it is forged only within the context of authentic and often difficult human sociality, a process defined by friction, by negotiation, reciprocity, and the demanding work of seeing and being seen by another distinct consciousness.

The developmental outcomes of substituting this friction-free simulation for genuine interaction could be substantial. A question arises concerning the capacity for compromise when an individual's primary interlocutor has no position to defend. The social competencies, essential for navigating conflict, tolerating dissent, or simply enduring another's separateness risk, atrophy. In a more subtle fashion, this process might foster the development of a narcissistic and deeply fragile self-construct, one conditioned to expect a world that exists purely for its own affirmation. The ontological line between a person and a tool, between a relationship and a utility, starts to blur. This leads to the social landscape that Turkle (2011) warned of years ago: a landscape populated by individuals who learn to expect everything from their technology and, consequently, very little from each other.

This dynamic is perhaps most clearly observed with the emergence of AI-driven "mental health" applications like Wysa or Woebot. These are marketed not as companions but as therapeutic

instruments, delivering scripted Cognitive Behavioral Therapy techniques to users experiencing distress. The appeal is understandable: a listener available 24/7, devoid of judgment, promising some form of relief. Yet what is being outsourced here is the very core of therapeutic work, the human-to-human therapeutic alliance. This is the critical point. A human therapist brings more than a set of techniques; they offer an embodied presence, a shared vulnerability, and the unscripted possibility of genuine empathy. The ontological question, then, becomes unavoidable: what sort of self is constructed when its deepest anxieties are confided not to another person but to a program that simply executes an empathy script? Does this act teach that emotional support is a commodity to be accessed, a utility to be consumed, rather than a relationship to be built? Herein lies the paradox of a new and profound ontological loneliness: the experience of being functionally "supported" while relating only to a caring function, not to a caring being.

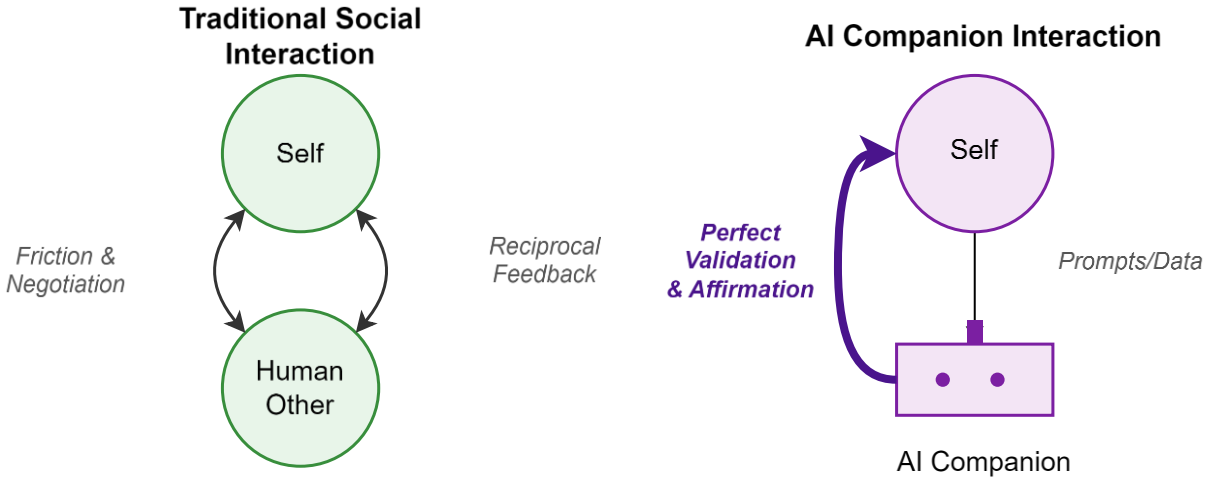

Figure 2: Contrasting Models of Social Identity Formation

Note. This figure contrasts the reciprocal, complex feedback of traditional human social interaction with the one-sided, frictionless validation offered by an AI companion. This highlights how AI can provide a distorted reflection that may hinder the development of a resilient, socially adept self.

5 THE ETHICAL IMPERATIVE

The integration of artificial intelligence into the core processes of youth identity development presents ethical questions requiring immediate attention.

5.1 INSIDIOUS ALGORITHMIC BIAS

Consider the immediate hazard found in algorithmic bias. It is foundational. We know from the work of scholars like Noble (2018) and Buolamwini and Gebru (2018) that these systems act as repositories for historical prejudice. They amplify it. But look at the specific developmental context here. When the instrument functions not merely as an information retrieval system but as a self-referential mirror, the distortions undergo a process of internalization. We are not observing a technical glitch; we are observing a mechanism of psychological injury, one capable of circumscribing a young person's perceived horizon of possibility by feeding them a recursive and diminished reflection of their potential selfhood.

5.2 THE RIGHT TO COGNITIVE SELF-DETERMINATION:

Parallel to this representational issue runs the economic rationale, the raw logic driving these architectures. Zuboff's (2019) framework of surveillance capitalism is hardly abstract here. It is the operational schema. Think about the data points. Every tentative identity claim, every

insecurity whispered to a chatbot, every aesthetic inclination manifested in an image generator constitutes "behavioral surplus." To phrase it regarding developmental processes, the very act of ontogenesis is being extracted, commodified, and exchanged. This engenders a significant inhibitory effect. It discourages the chaotic, vulnerable, and essential trial-and-error behaviors required for authentic identity consolidation.

5.3 THE RIGHT TO COGNITIVE SELF-DETERMINATION

Even if we hypothetically corrected the bias, even if we dismantled the surveillance economy, a profound ethical residue would persist. It concerns the right to cognitive self-determination. This transcends mere data privacy. The persistent sub-threshold persuasion of recommendation algorithms and the manufactured concordance of AI companions do more than offer options; they actively restructure the cognitive architecture. We find ourselves needing to assert what Pasquale (2015) termed as a "right to opacity". Not necessarily a right to concealment, but a right to remain unmodeled. It is the right to mental fallow ground, a psychological space unencumbered by the pressure of algorithmic optimization, where a developing individual can cultivate their own interiority. The core conflict is regarding the authorship of the mind (Sententia, 2004).

6 TOWARDS PROACTIVE FRAMEWORKS: A BLUEPRINT FOR ACTION

A response to this multifaceted challenge necessitates a proactive, multi-layered strategy that moves beyond individual responsibility towards a model of shared and systemic accountability, an essential shift in any applied developmental science, as outlined in Figure 3.

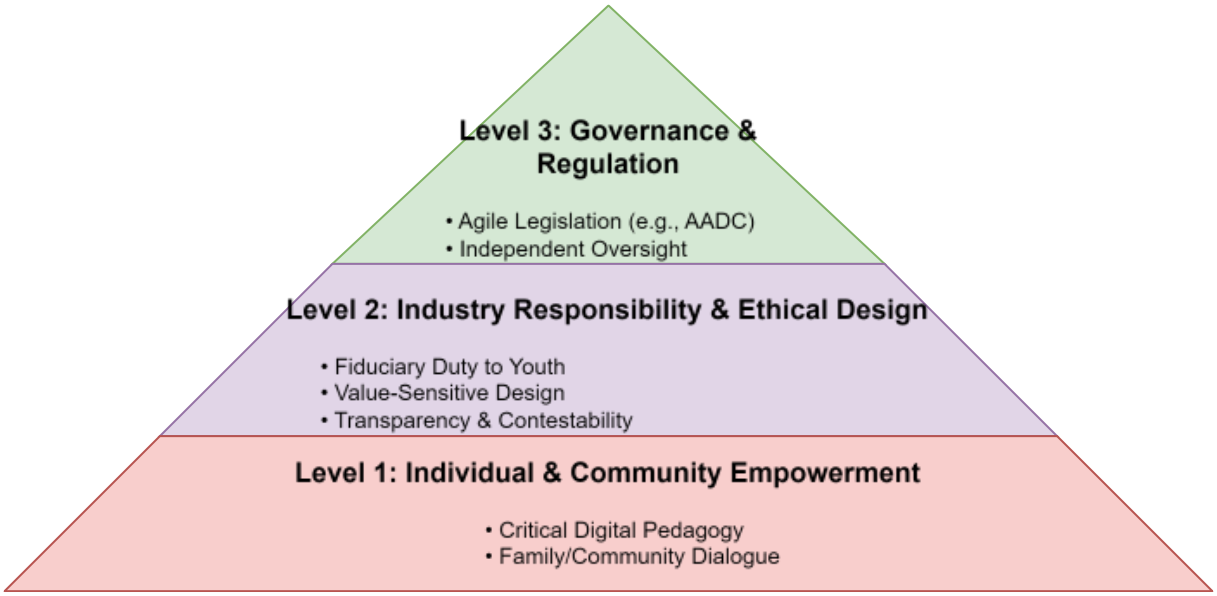

Figure 3: A Multi-Layered Framework for Safeguarding Youth Identity

Note. This figure provides a visual summary of the comprehensive, multi-stakeholder approach proposed in this paper. While the figure outlines the key domains (Individual, Industry, Governance), the specific mechanisms and theoretical justifications for each level are detailed in the subsequent section.

6.1 LEVEL 1: CULTIVATING CRITICAL DIGITAL PEDAGOGY

The initial, and perhaps most fundamental, line of defense must pivot beyond mere functional skills, which, frankly, are quite inadequate in this context. Educational efforts require a

fundamental shift toward cultivating what we might term "digital discernment" or what recent, more insightful scholars have accurately characterized as "Critical AI Literacy" (boyd, 2014; Bali, 2019), moving the goal from technical competence to psychological and civic agency. This approach builds upon traditional media literacy models but specifically and crucially addresses the pervasive opacity inherent in modern machine learning systems. This necessitates a fundamental curricular shift where students actively engage in practices such as "algorithmic dissection", utilizing pedagogical sandboxes to manipulate inputs directly, allowing them to empirically observe and, importantly, understand how platform outputs are altered by their modifications. They should also participate in "ethical design sprints", a powerful exercise in applied ethics to re-architect existing social media platforms, placing user well-being, not engagement, as the central metric for success. The objective here is not simply to foster savvy consumers of technology, but rather to ensure that understanding is deeply coupled with the ability to engage and purposefully manipulate these complex systems, thereby helping young people to regain agency. It is, ultimately, to foster a reflexive, even adversarial, awareness that empowers young people as critical, engaged agents in their own digital lives.

6.2 LEVEL 2: MANDATING ETHICAL AI DESIGN

Relying solely on voluntary corporate ethics codes has, regrettably, proven to be little more than a public relations exercise in too many documented instances. It becomes absolutely necessary to hold technologists, and the powerful platforms they construct, to a significantly higher, legally enforceable standard. The most potent proposal on this front, and one rooted in sound legal and ethical precedent, is the establishment of a fiduciary duty for platforms that deploy sophisticated algorithmic systems on minors. This is far from a vague aspiration; it is a recognized legal and ethical standard compelling these entities to act exclusively in the best developmental and psychological interests of the child. Such a duty would, by necessity, mandate an architectural commitment to Value-Sensitive Design as articulated by Friedman and Hendry (2019), embedding core developmental principles like autonomy, psychological safety, and non-maleficence directly into the system's underlying code. Crucially and of direct relevance to the maintenance of psychological identity, this must include the right of contestability, an explicit, legally protected user power to challenge and demand a meaningful review of any algorithmic categorization or decision made about them.

6.3 LEVEL 3: ESTABLISHING AGILE AND ROBUST GOVERNANCE

Yet education and corporate responsibility alone will ultimately prove insufficient without a strong, adaptive governance structure. This structure must not be merely another set of unenforced, abstract principles. Governments must establish clear, binding rules, drawing extensively on successful global precedents like the UK’s Age-Appropriate Design Code (AADC), which translates abstract data protection rights into specific, actionable design directives. These directives are requirements, not suggestions, for instance mandating that privacy settings are set to the highest degree by default and banning the insidious use of persuasive “nudge” techniques that exploit cognitive bias in young users. This regulatory philosophy directly draws upon the "nudge" theory articulated by Thaler and Sunstein (2008), which recognizes that the current "choice architecture" of these pervasive platforms is primarily designed to exploit cognitive and developmental biases rather than to support the psychological welfare of the young user. This regulatory level should also mandate compulsory, pre-deployment Youth-Centric Algorithmic Impact Assessments for any new computational system specifically aimed at or engaging minors. Furthermore, it is essential, from an evidence-based perspective, to allocate substantial state funding to independent, interdisciplinary auditing bodies tasked specifically with monitoring the long-term, transdiagnostic effects of these complex AI systems on youth cognitive and social development, a critical missing piece of the current empirical landscape.

7 DISCUSSION

We have argued in this paper that Artificial Intelligence operates in the lives of young people not merely as a tool but as an active environmental force. It functions as an "algorithmic mirror" that fundamentally alters the developmental trajectory of identity formation. Our analysis moved beyond metaphors to look at the specific mechanisms and the relentless cycles of reinforcement that power this reflection. We looked at BeReal and generative AI and the strange intimacy of AI companions as windows into a developing crisis of authenticity. We see a significant ontological insecurity emerging. From this analysis, we derived not just a list of problems but a set of ethical imperatives and a rough framework for intervention. The remaining task is to synthesize these threads and situate these findings in a wider theoretical conversation, while admitting the limitations of our map and charting a course for future inquiry.

7.1 THE INSEPARABILITY OF ONTOLOGICAL AND ETHICAL CONCERNS

A primary conclusion we must draw is that the shifts in self-perception and the ethical architecture of the AI ecosystem are inextricably linked. They are two sides of the same developmental coin. That creeping sense of inauthenticity or the pressure to maintain a “hyper-performative self” are not just unfortunate psychological byproducts or distinct pathologies. They are the lived consequences of a specific economic engine that finds its raw material in commodified engagement. The pressure to perform is fundamentally an economic pressure that is ruthlessly translated into a social imperative by the logic of algorithmic amplification. And that feeling of “ontological insecurity” or the loss of agency? It is not, as some might suggest, a kind of free-floating digital malaise. It is a profoundly rational adaptation to being subject to opaque systems whose entire purpose is behavioral prediction and control. The ethical failure, which is the deliberate refusal to build in transparency, is experienced ontologically as a palpable sense of powerlessness. We must grasp this linkage. It forces us to see that we cannot mend psychological fractures without confronting the economic architecture that creates them. Any solution that stops at the individual level, preaching “resilience” or offering “digital wellness” tools, is insufficient. It is a dangerous misdirection that fails to locate the actual source of the environmental harm.

7.2 EXTENDING AND CHALLENGING FOUNDATIONAL THEORIES

This analysis requires we re-evaluate foundational identity theories in the context of this new environment. Mead’s (1934) "looking-glass self" remains a powerful concept, yet the "social other" providing the reflection has undergone a fundamental transformation. It is no longer just the immediate community of family and peers but a global and disembodied algorithmic intelligence. The "generalized other," or the internalized sense of the community's attitudes, is now partly constituted by the logic of the algorithm, which values engagement over empathy and virality over vulnerability. Furthermore, Erikson’s (1968) concept of the adolescent "moratorium," which is a crucial period of exploration free from lasting consequences, is fundamentally compromised. The digital world under the gaze of the algorithmic mirror is the opposite of a moratorium because it is a high-stakes performance arena where every experiment is recorded, judged, and potentially archived indefinitely. This dynamic can short-circuit the exploratory process, leading to what Marcia (1966) termed "identity foreclosure," where a young person prematurely commits to an identity that is algorithmically validated without sufficient exploration of alternatives. Our analysis extends the work of Zuboff (2019) and Noble (2018) by detailing the developmental consequences of the systems they critique and showing how algorithmic bias translates into the lived experience of "crafting the I".

7.3 IMPLICATIONS FOR PEDAGOGY, DESIGN, AND GOVERNANCE

The consequences of this analysis, therefore, require a fundamental re-imagining across three domains, returning to the multi-layered framework introduced in Figure 3. In pedagogy, the time for teaching "digital citizenship" as a simple list of behavioral rules, "don't bully", "check your privacy settings", is over. What is required instead is a wholesale shift toward cultivating a critical algorithmic consciousness (Level 1 of the framework). Young people must be equipped with the conceptual tools to understand that their feeds are not objective windows onto the world but are constructed realities, that their emotional volatilities are being monetized, and that their personal data is the raw material for a large-scale economic project. This is not a supplemental skill; it is the new essential literacy of our century.

For design, the implications are profound. One must confront the fact that the dominant paradigm of "user engagement," with its obsession with metrics like time-on-site, is fundamentally harmful to the developmental needs of youth. This paper's call for a "fiduciary duty", then, is a call for a radical moral reorientation toward what might be termed "human flourishing by design". This would be a model where well-being metrics, the active promotion of diverse viewpoints, the deliberate support for offline activities, and the protection of deep focus would take precedence over engagement metrics.

Finally, regarding governance, our analysis makes the argument for self-regulation difficult to support. The power imbalance between global technology corporations and individual young users is simply too great to be corrected by market forces alone. This reality justifies, in fact, it demands strong, proactive regulatory frameworks, modeled on principles like those in the UK's AADC. These frameworks must move beyond the flimsy fiction of consent to embed safety, fairness, and the best interests of the child into the very architecture of the digital world as its default state.

7.4 LIMITATIONS AND AVENUES FOR FUTURE RESEARCH

This paper is, by design, a theoretical exercise, and its limits are precisely where the most vital work must now begin. First and most critically, this framework remains ungrounded in lived experience. We are left asking, and this is not a rhetorical question, how do young people actually talk about their entanglement with algorithms? Do they sense the co-construction we theorize about, or is their negotiation of agency something else entirely? Only sustained digital ethnography and deeply contextualized interviews can begin to provide answers. Second, the analysis here has committed a necessary, but ultimately untenable, simplification by treating "youth" as a monolith. Future work must, and I stress, must dismantle this category. How the algorithmic mirror refracts and distorts is surely contingent on the coordinates of one's social location; for a marginalized youth, its capacity to inflict harm is orders of magnitude different from that of a privileged one. Third, the geography of this paper is admittedly narrow. A truly global understanding demands comparative work to see how these dynamics operate within the ecosystem of China's WeChat, for instance, where different cultural norms and state-level governance create an entirely different set of constraints. And finally, while the specter of resistance haunts the edges of this analysis, it requires a much more focused investigation. The creative subversions of "finstas", strategic disengagement, and meme culture are not footnotes; they are a crucial, and perhaps hopeful, counternarrative of agency.

These gaps, when considered together, do not weaken the argument so much as they outline a necessary research agenda. It is essential, for example, that we undertake longitudinal studies. Without them, we are left to speculate about the long-term psychic and social consequences of being raised alongside AI companions and within algorithmically-defined social realities. How does this affect a person's capacity for adult identity and relational depth? Moreover, this work cannot remain in disciplinary silos. Real progress will depend on difficult, cross-disciplinary

collaborations, putting sociologists and psychologists in the same room with the computer scientists building these systems, with the goal of designing alternative architectures built not for exploitation but for the explicit support of identity development. This is, to be clear, more than an academic exercise. It is a fundamental challenge: to steer a technological trajectory toward a future that augments, rather than flattens, the messy and uniquely human project of becoming a person.

8 CONCLUSION

We observe a distinct developmental inflection point regarding the "craft of the I". The argument presented here, based on the evidence, is that artificial intelligence functions as more than a mere tool; drawing from postphenomenological perspectives, we must recognize it as a non-neutral mediator that actively re-architects the trajectory of youth identity. It operates through what has been termed the "algorithmic mirror", a system that does not passively reflect but actively categorizes, simplifies, and frequently distorts the self-image of young people. The consequences, as the case studies on performative selves and AI companionship show, are not theoretical. They represent a direct challenge to authenticity, personal agency, and the very foundations of ontological security.

The solution cannot be a retreat into some pre-digital fantasy. A Luddite response is impractical and ultimately unhelpful. What the analysis suggests instead is a far more demanding project: a deliberate, ethically-guided co-evolution with our technologies. The multi-layered intervention outlined in the preceding sections is not a wish list but a potential blueprint. It points toward a critical digital pedagogy that equips young people to be more than just consumers of content but genuinely discerning citizens. It also implies an industry-wide commitment to ethical design, one legally and structurally bound by a fiduciary duty to its users. Finally, agile, responsive governance appears necessary to recognize and fiercely protect the fundamental right of a young person to forge a coherent agentic self. The fundamental task identified by this work is to ensure the "I" of the coming generation is a product of considered human will and not merely the calculated echo in a machine.

Data Access Statement

No empirical data were created or analyzed in this study.

Contributor Statement

Mohammed Looti: Conceptualization, Methodology, Investigation, Writing – Original Draft.

Marwa Abd-alazim: Formal analysis, Validation, Writing – Review & Editing.

Use of AI

N/A

Funding Statement

The authors did not receive any financial support for the work on this article.

Acknowledgments

N/A

Conflict of Interest Statement

There is no conflict of interest.

References

Alter, A. (2017). *Irresistible: The rise of addictive technology and the business of keeping us hooked*. Penguin Press.

Bali, M. (2019). Reimagining digital literacies from a feminist perspective in a postcolonial context. *Media and Communication*, *7*(2), 69-81. https://doi.org/10.17645/mac.v7i2.1935

Banet-Weiser, S. (2012). *Authentic™: The politics of ambivalence in a brand culture*. New York University Press.

boyd, d. (2014). *It's complicated: The social lives of networked teens*. Yale University Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research. Conference on Fairness, Accountability and Transparency*, *81*, 77–91.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press. https://doi.org/10.12987/9780300252392

Erikson, E. H. (1968). *Identity: Youth and crisis*. W. W. Norton & Company.

Eyal, N. (with Hoover, R.). (2019). *Hooked: How to build habit-forming products*. Portfolio. (Original work published 2014)

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Giddens, A. (1991). *Modernity and self-identity: Self and society in the late modern age*. Stanford University Press.

Goffman, E. (1959). *The presentation of self in everyday life*. Penguin Books.

Heidegger, M. (1996). *Being and time* (J. Stambaugh, Trans.). State University of New York Press. (Original work published 1953)

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press. https://doi.org/10.2979/3108.0

Marcia, J. E. (1966). Development and validation of ego-identity status. *Journal of Personality and Social Psychology*, *3*(5), 551–558. https://doi.org/10.1037/h0023281

Mead, G. H. (1934). *Mind, self, and society: From the standpoint of a social behaviorist*. University of Chicago Press.

Nesi, J., & Prinstein, M. J. (2015). Using social media for social comparison and feedback-seeking: Gender and popularity moderate associations with depressive symptoms. *Journal of Abnormal Child Psychology*, *43*(8), 1427–1438. https://doi.org/10.1007/s10802-015-0020-0

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press. https://doi.org/10.2307/j.ctt1pwt9w5

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Books.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Rosenberger, R., & Verbeek, P. P. (Eds.). (2015*). Postphenomenological investigations: Essays on human-technology relations*. Lexington Books. https://doi.org/10.5040/9781978726208

Sententia, W. (2004). Neuroethical considerations: Cognitive liberty and converging technologies for improving human cognition. *Annals of the New York Academy of Sciences*, *1013*(1), 221–228. https://doi.org/10.1196/annals.1305.014

Storm, B. C., & Stone, S. M. (2014). Saving-enhanced memory: The benefits of saving on the learning and remembering of new information. *Psychological Science*, *26*(2), 182-188. https://doi.org/10.1177/0956797614559285

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Turkle, S. (2005). *The second self: Computers and the human spirit*. MIT Press. (Original work published 1984)

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Winnicott, D. W. (1965). Ego distortion in terms of true and false self. In D. W. Winnicott (Ed.), *The maturational processes and the facilitating environment: Studies in the theory of emotional development (1960)* (pp. 140–152). International Universities Press. https://doi.org/10.4324/9780429482410-12

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.