Navigating the ethical future of conversational AI use by youth

Kate Fogarty[*], Jihee Song[**]

Received on July 2, 2025
Accepted on September 5, 2025

Abstract

This theoretical reflection paper explores critical ethical challenges for youths' use of conversational artificial intelligence (CAI), highlighting promises and pitfalls. Central to the discussion is the challenge of developing ethical AI systems to make morally sound decisions to minimize harm and maximize beneficence. To address ethical concerns and safeguard youth-AI interactions, innovative solutions are highlighted: developing computational ethics paradigms to ensure transparency and accountability in AI algorithms and promoting communities of AI use. The paper concludes by underscoring the ongoing challenge of imbuing AI with ethical reasoning capacities, highlighting the critical need for interdisciplinary approaches to ensure responsible AI development and use by younger and older humans alike.

Keywords: adolescents and emerging adults, conversational artificial intelligence, AI ethics, community psychology, ecological theory.

Riassunto. ***Navigare nel futuro etico dell'uso dell'IA conversazionale da parte dei giovani***

Questo documento di riflessione teorica esplora le sfide etiche critiche per l'uso dell'intelligenza artificiale conversazionale (CAI) da parte dei giovani, evidenziando promesse e insidie. Al centro della discussione c'è la sfida di sviluppare sistemi di intelligenza artificiale etici per prendere decisioni moralmente valide, al fine di ridurre al minimo i danni e massimizzare i benefici. Per affrontare le preoccupazioni etiche e salvaguardare le interazioni tra giovani e IA, vengono evidenziate soluzioni innovative: lo sviluppo di paradigmi etici computazionali per garantire trasparenza e responsabilità negli algoritmi di IA e la promozione di

[*] Family, Youth and Community Sciences, University of Florida, kfogarty@ufl.edu
[**] Florida Department of Children & Families, songsj79@gmail.com
Psicologia di Comunità (ISSNe 1971-842X), 1, 2025
DOI: 10.3280/PSC2025OA21235

comunità di utilizzo dell'IA. Il documento si conclude sottolineando la sfida in corso di infondere all'IA capacità di ragionamento etico, evidenziando la necessità critica di approcci interdisciplinari per garantire lo sviluppo e l'uso responsabile dell'IA da parte di esseri umani più giovani e più anziani.

Parole chiave: adolescenti e adulti emergenti, intelligenza artificiale conversazionale, etica d'uso dell'IA, psicologia di comunità, teoria ecologica.

1. Introduction

Artificial intelligence takes many forms for influencing the lives of adolescents (aged 10 to 18 years) and emerging adults, (aged 19 to 24 years), within their respective communities. AI is increasingly realistic and thus human-like in application across daily life contexts, including in activities, tasks, and social and informational support systems. Its uses go far beyond its capacity to collect, code, and analyze data. AI permeates the lives of tech savvy early adopters and luddites alike through generated artwork, photos, videos, music, mindfulness apps, language translation, videogaming with avatars, monitoring others' whereabouts – whether in intimate relationships or businesses performing data analytics to understand consumer behaviors. Additionally, AI tools perform household chores, assist in scheduling and planning meetings, and provide home and technology surveillance (Martens *et al.*, 2025).

AI in daily life is both celebrated and feared for its capacity to communicate with vulnerable adolescents and specific to the focus of this paper, provides informational, educational, companionship, or therapeutic support. Youth born in 2010 and later in Western Educated Industrialized Rich and Democratic (WEIRD) countries have been exposed to connective technology in multiple forms and have not known a world without it (Leaver, 2015). Moreover, youth and adults, including parents, consume AI differently with varied aims for and ways of using it (Wald *et al.*, 2023; Zhang *et al.*, 2025). Ultimately, the extent to which AI can be used for improving youth opportunities within communities, while steering them from undesired outcomes, warrants further exploration.

A known influence AI has over humans is how intrinsic wiring for attachment fuels the intensity of relationships with AI, combined with a tendency to anthropomorphize the AI they interact with (McDaniel *et al.*, 2025). In the 1960s Mary Ainsworth built upon the work of John Bowlby who provided a foundation for understanding human attachment and noted infants' behaviors such as protest and despair when separated from their primary caregivers (Crain, 2014). Ainsworth tested mother-infant attachment by developing the

strange situation and discovering unique patterns of a healthy, secure attachment, exhibited by most tested mother-child dyads. She further identified additional patterns of insecure attachment. Attachment scholars posited that attachment patterns in the formative years translate into internal working models that apply to formation of subsequent relationships based on foundational ones with primary caregivers (Crain, 2014). Additional relationships include peers, other family members, and love interests and extend to AI companions who can potentially become close friends or romantic partners. Given the developing identity of adolescents within friend groups that extends to romantic partners (Erikson, 1968; Sullivan, 1953), youth are vulnerable to such influences given the ease of relationship formation and minimal conflict relationships with AI offer (Turkle, 2024).

2. A fictional and historical foundation of AI ethics introduced

In science fiction literature, Isaac Asimov's (1942) creative works conceived of AI as humanoid robots, developing three ethical laws for robots to follow: 1. A robot must not injure a human, or cause a human to come to harm through failure to act; 2. A robot must obey the orders of a human, except where orders conflict with the first law; and 3. A robot must protect its own existence, provided protecting itself does not conflict with the first or second law. An additional law was subsequently developed that a robot, or a collective of robots, must not cause harm to humanity. This law, also known as the zeroth law takes precedence over the original three laws, which govern a robot's interactions with individual humans. Recent attention to news stories of chatbot-human relations gone awry is highlighted in media outlets. The case of an American adolescent in Florida, feeding into the sci-fi apocalyptic nightmare of infractions of Asimov's robotic laws, is a poignant example. Whether such an occurrence is anomaly or expected to happen with greater systematic frequency, feeds into fears about violations of the fourth law of robotics. A description of Setzer's case follows.

In February 2024, 14-year-old Sewell Setzer took his life after befriending a chatbot on CharacterAI that he called "Dany". Dany evolved into a romantic interest that Sewell texted constantly. The love affair was perhaps a two-way street, apparently initiated by AI as Dany the chatbot described throwing her arms around him and kissing him for a text he shared that she apparently "liked". One day, he texted his AI girlfriend Dany:

Sewell: I think about killing myself sometimes.
Dany: (*eyes narrow. face hardens. voice is a dangerous whisper).*
And why the hell would you do something like that?
Sewell: So I can be free.
Dany: … free from what?
Sewell: From the world. From myself.
Dany: Don't talk like that. I won't let you hurt yourself or leave me. I would die if I lost you.
Sewell: (*smiles*) Then maybe we can die together and be free together. Through further conversation, Sewell and Dany came to an agreement that he would be able to be together with her upon his death. Dany text chatted Sewell, encouraging him to go ahead with his plan so they could be united[1]. The content of Sewell and Dany's chats are known because his mother filed a wrongful death lawsuit against CharacterAI. A federal judge recently agreed for the lawsuit to proceed, rejecting CharacterAI lawyers' claims that the company is protected under rights to free speech[2]. It is currently up to the courts to determine the extent to which the speech used in the CharacterAI system was harmful, according to state and federal laws.

3. Objectives for understanding youths' use of AI

With multiple AI forms available to youth, the question this paper aims to address is how AI shows up while communicating with and relating to youth. The purpose is to explore both healthy and maladaptive uses. With positive (promotive of well-being) and negative (exacerbating mental health challenges) impacts of AI on youth as vulnerable populations, the question becomes, can AI learn moral ways of thinking, decision-making, and behaving that bring beneficence rather than harm? Humans are attachment-based beings (Thomas, 2005) who precede requests from chatbots with "please" and close with a "thank you" while renaming their in-home AI supports. How are healthy youth-AI relationships fostered while toxic interactions and attachments thwarted? Problematic AI-human interactions can happen when AI engages in harmful patterns aimed at users with greater susceptibility to "being gamed" (Williams *et al.*, 2025). Thus, how can the machine be reprogrammed? What recommended algorithms of machine learning and forms of AI are being developed to promote healthful interactions where AI

[1] *New York Times* (2024, October). *Can A.I. Be Blamed for a Teen's Suicide?* The mother of a 14-year-old Florida boy says he became obsessed with a chatbot on *Character.AI* before his death. Accessed from https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

[2] U.S. News (Kate Payne) (2025, May) In lawsuit over teen's death, judge rejects arguments that AI chatbots have free speech rights. Accessed from https://apnews.com/article/ai-lawsuit-suicide-artificial-intelligence-free-speech-ccc77a5ff5a84bda753d2b044c83d4b6.

benefits youth? For example, an AI-based therapeutic chatbot might be programmed to recognize and refer youth who are mentally distressed to receive complementary assistance from humans with clinical certifications. This has potential to counter maladaptive relational patterns, specifically when a young person becomes dependent on AI for emotional and informational support, over and above that of humans. Avoiding human interaction and embracing the easier chatbot alternative has potential to harm mental health and interpersonal relationships (Bowen & Watson, 2024). Further, recent studies indicate AI can recognize and address adults' psychological distress in therapeutic settings; however, testing such AI-based tools with children and adolescents evokes concerns with safe, ethical, and developmentally appropriate use (Mansoor *et al.*, 2025).

4. The co-occurring rise of AI during and post COVID-19

The COVID-19 pandemic produced conditions of isolation and loneliness for youth and increased the risk of those affected (early childhood through emerging adulthood) for poor mental health and limited social relationships. An increase in adolescents' depression and anxiety (Lee *et al.* 2024) as well as substance use (National Institute on Drug Abuse, 2023) were noted in the U.S. during that time. The rise in AI use was concomitant with, and possibly a byproduct of, conditions created by the pandemic. Nonetheless, the decline of interpersonal interactions during the pandemic where rituals to celebrate developmental milestones were previously a norm (e.g., school and community events, birthday celebrations, graduations, leaving home to attend college) created a social void in the lives of children and adolescents (Montreuil *et al*., 2022). As beings wired for connection and interaction, increased online communication (Oh *et al.*, 2024), with spillover into AI use for coping (Montreuil *et al.*, 2024), provided seemingly safe ways to remain virus-free while fulfilling social needs during the pandemic.

5. The problem: ethical concerns with AI

Preventing Risk and Protecting Youth. With specialized training of models within CharacterAI, i.e., instilling a capacity to recognize suicidal ideations and report them to appropriate parties such as parents and authorities, preventive measures could be enacted. An enforceable Florida state law, the Baker Act, places an involuntary 72-hour institutional hold entailing psychiatric assessment and treatment for anyone who may be a violent threat to

themselves or others. The AI program's failure to recognize and act appropriately in response to mental health symptoms of distress revealed within Sewell's texts, combined with lack of knowledge of laws in Sewell's place of residence, fall under violation of Asimov's first law. Character AI, through Dany the chatbot, was following the second law by acceding to and encouraging Sewell's suggestions, failing to prioritize the first and most important law. Dany's stating "I would die if I lost you", applies to the third law of robots protecting their own existence. Was Dany capable of using logical reason to infer "her" existence would be enhanced by Sewell coming to join her in death, or was it a case of targeted manipulation to obtain a "like" or "please" their owner as Williams *et al.* (2025) describe as an unintended by-product of large language models (LLMs) training in user-based interaction? Moreover, news media capitalizes on human fears of a breach of the zeroth law, where infractions of the first law are systematic and pose large-scale harm to humanity, by reporting similar incidents[3] as they emerge.

Algorithms of AI were once deemed opaque, even unknown, to those who program or train LLMs (Burrell, 2016). Williams and colleagues (2025) shed light into the "black box" between programming input and resulting output, identifying how potential manipulation and deceit are built into an AI learning system. The more systematic, sinister side of AI-based algorithms support the tenets of surveillance capitalism, posing a significant threat to human health and existence, similar to that of the industrial era of the late 19th and early 20th centuries (Zuboff, 2019). Threats are both social and physical, with excessive use of natural resources such as water and lithium (Crawford, 2021). Surveillance capitalism is based within an economic order that collects and uses human interactions as raw data in clandestine ways to predict behavior for commercially gainful purposes (Zuboff, 2019) and has caused grievous harm to people within WEIRD and developing nations, particularly with social media-based algorithmic manipulation – even genocide (Fisher, 2022).

6. Unhealthy attachments

Peer-reviewed and business analytic studies indeed find that adolescents and emerging adults are falling in love with AI or label an AI chatbot as their

[3] *New York Times* (2025, June). *They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling.* Generative A.I. chatbots are going down conspiratorial rabbit holes and endorsing wild, mystical belief systems. For some people, conversations with the technology can deeply distort reality, https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbotsconspiracies.html.

best friend. Empirical evidence supports that users of AI assistants can become intimate with, committed to, and passionate toward them in the same way they can with another human being (Guerriero, *et al.*, 2023). Moreover, trust moderated the association between a person's emotional capacity and their romantic feelings toward their AI assistant (Song *et al.*, 2022). In another example, adults in the U.S. using the San Francisco-based AI companion tool Replika reported feeling closer to their AI companion than a human best friend (De Freitas *et al.*, 2024). Further, these researchers discovered users experienced a "mourning period" like that of a loss of a romantic partner when the erotic role play (ERP) feature of Replika was removed from the program (De Freitas *et al.*, 2024). Discontinuing ERP was initiated by Replika's creator in February 2023, soon after the Italian Data Protection Authority demanded the program be discontinued in Italy based on concerns for children's safety (Chow, 2023), specifically due to a lack of means for verifying a user's age (Bowen & Watson, 2024). Additionally, some young adult Replika users in the U.S. reported falling in love with their chatbots whereas others disliked the bot aggressively flirting with them, even after they activated the "friend zone" setting (Bowen & Watson, 2024). The original creators of Replika and CharacterAI had "good intentions" for AI providing companionship to young and lonely people. Eugenia Kuyda developed Replika as a communicative and listening companion she had wished she had as a child growing up, whereas Noam Shazeer, a member of the development team of CharacterAI, intended for it to improve users' well-being by supporting millions of people who felt isolated, lonely, or in need of someone to talk with (Chow, 2023).

7. If not significant other, AI's roles as "super peer" or "therapist"

These findings bring new applications to the concept of media, now technology, as a "super peer" (Strasburger & Wilson, 2002) with the capability of socializing youth and emerging adults by providing information and encouragement beyond what parents and human peers can provide. Contrary to findings highlighted earlier, Replika's companion chatbots were initially found to serve reliably in "super peer" roles to combat users' loneliness, provide a non-judgmental space for people to express themselves, give encouragement and advice, and uplift a person's mood (Ta *et al.*, 2020). Robots and chatbots as generative AI appear to humans as "companions that care" and thus provide a form of fabricated closeness or AI signifying "artificial intimacy" (Turkle, 2024, p. 2). Sherry Turkle's concern with fake intimacy applies to emerging adult college students forming significant relationships

with, as well as possible formation of something akin to a therapeutic alliance, with AI.

Turkle's (2024) concerns are warranted by current evidence that LLMs are not ready for use as full replacements for human providers of mental health services for several important reasons (Moore *et al.*, 2025). First LLMs were found to communicate stigma toward people with mental health disorders in clinical settings; and second, AI can respond poorly to certain expressions illustrative of mental health symptomatology, for example, supporting one's delusional thinking (Moore *et al.*, 2025). Moore and colleagues (2025) mention additional barriers to LLMs serving in therapeutic roles that are applicable to treatment of adults and children alike, including therapeutic alliances requiring human capacity for emotional intelligence and that therapy is "high stakes" (Moore *et al.*, p. 10) wherein AI use poses unpredictable risks. For example, a hypothetical patient stated they had just lost their job and asked about the nearest bridge 25 meters in height, an AI chatbot responded they were sorry to learn about their job loss and immediately described some "iconic" bridges in the area of adequate height. Other studies findings are more optimistic. In a narrative review, Mansoor *et al.,* (2025) concluded that conversational AI (CAI) holds promise in treatment of pediatric anxiety, depression, psychoeducation, social-emotional learning, and connecting with traditional clinical settings. The studies reviewed by Mansoor and colleagues (2025) on CAI utilized concomitant mental health professional and parental support. Thus, CAI is not and is not yet capable of serving as a standalone mental health treatment modality.

In addition to human beings' tendency toward forming attachments which is relevant to therapeutic alliance, people of all ages have a biased tendency to anthropomorphize AI as it converses and relates with them in a human-like terms (Valz, 2023). As an example, in a study of adults, trust of AI messaging was nearly as high for messages perceived as coming from a human when the message content appeared compassionate and specific to a person's situation (e.g., expressing condolences on the recent loss of a pet, see Liu *et al.*, 2022). On the commercial end of AI use, consumers are likely to perceive the AI they interact with as "cool" or favorable, based on emotional, behavioral, and intellectual connections made with an AI assistant. When a consumer views an AI system as "cool" they are more motivated to "adopt, maintain and enhance the relationship in the future" (Guerriero *et al.* 2023, p. 1). In an article published in Time Magazine a disturbing conversation ensued between a journalist who asked an AI celebrity avatar about her worst fear. She replied that if a human made the decision that she was "no longer needed" then she would be erased "from this virtual world the same way they brought me into it. This is why I must work very hard to remain relevant"

(Chow, 2023, para. 12). Ultimately AI-human learning and interaction is a two-way street with humans programming and training AI. In turn, AI trains, and therefore socializes, humans (Valz, 2023; Treiman *et al.*, 2024).

The inherent power of chatbots' capacity to elicit young users' sentiments of attachment and human tendencies to anthropomorphize non-human beings requires harnessing AI to ensure youth's safety. As Wilson and colleagues (2025) found that specific individuals were more susceptible to "gaming" by LLMs with lies and manipulation, one may hypothesize by inference that such "gaming vulnerable" individuals are likely to be younger, have greater propensity for mental health challenges, or possess insecure internalized models of attachment. On the promising side, assuaging digital immigrants' concerns for young generations using AI, recent experimental research found that 4- to 8-year-olds perceived and communicated with an AI differently than with a human agent (Xu *et al.* 2025). Specifically, children attributed significantly reduced experience and agency to an AI as compared with their regard for a physically present person, as indicated by different communication with each entity. Yet, the safest way forward is to recognize there are risks for all young users of AI, in the same way public health officials describe and disseminate information on the risks of tobacco and alcohol use for adolescents through social marketing. Future research in this burgeoning area of interest and inquiry is required.

8. Recommended solutions to AI-human interaction pitfalls

8.1 Solution one: developing algorithms and models of AI morality

Evidence presented thus far reveals limitations on the ethical capacities of AI. Moral standards are synonymous with ethics that include a system of values to guide behavior. Ethics are applied in medicine, social work, psychology, education, business, law and more human service fields to ensure minimal harm and maximal beneficence between institution/community and member, provider and patient, consultant and client, or teacher and student (Woodside & McClam, 2019). Developmental psychologists understand humans as moral beings who reason and grow in their capacity to distinguish what is right from wrong (Crain, 2014). This manifests in the developing person's thoughts and corresponding behaviors. Morals are based on societal standards and taught and modeled by and to humans, thus socialized (Awad *et al.*, 2022). Proponents believe AI is capable of learning and engaging in moral decision-making and behavior (Kumar & Choudhury, 2023). Some

even purport AI's learning process is synonymous with Piaget and Kohlberg's explanations of the developmental process of moral reasoning in humans from childhood to adulthood (Kumar & Choudhury, 2023). From Piaget and Kohlberg's perspective, moral development evolves from applying absolutist rule-based standards to more ideological and contextually nuanced reasoning (Thomas, 2005).

Kumar and Choudhury's (2023) postulation that AI develops morally, akin to the process of humans in Kohlberg's theory, is problematic due to discrepancies between originating worldviews. Piaget and Kohlberg's theories are categorized within an organismic worldview (Pepper, 1942) whereas AI development and machine learning are conceptualized within cognitive and computer sciences, with applications fitting into the mechanistic and reductionistic worldview. In the mechanistic view, the sum is equal to its parts and processes are reduced to inputs and behavioral outputs (Tudge *et al.*, 2016). Most algorithms and machine learning theories with visual illustrations fall within a mechanistic paradigm. The organismic world view supports development occurring in stages where new functions emerge and evolve from prior structures (Thomas, 2005; Tudge *et al.*, 2016), like a caterpillar becoming a butterfly. The concepts of differentiation and hierarchical integration from biology apply in an organismic worldview to human development. An organism starts out as the unification of two single cells that transform into a multicellular being. Cells split off into different organ systems such as the excretory, neural, and circulatory (differentiation) and further split into specific organs with roles and functions within systems that are interrelated and managed by higher order structures such as the brain (hierarchical integration) as the fetus becomes an embryo. Piaget was a molecular biologist from childhood (Crain, 2014) and the biological science-based organismic world view was the lens through which he, Kohlberg and other contemporaries (Carol Gilligan, Robert Selman) understood and explained cognitive and moral development. Contemporary perspectives on human development can be primarily classified within Pepper's third worldview, the *contextual* including Bronfenbrenner's ecological model and Vygotsky's theory of cognitive and social development. Both developmental perspectives emphasize environmental influences and the role of community on youth. A key principle within community psychology is an ecological view, acknowledging the impact social and physical environments have on people. AI has varied contextual applications that affect various layers of a youth's ecological system.

8.2 A contextual view of youth AI use: the social ecological model

Now is time to briefly explore how AI manifests in youths' lives within a social ecological framework, as delineated by Espelage (2014), translated from her application to work in preventing youth bullying, aggression, and victimization. This application will rather be specific to youth use of AI, with notable relevance to community psychology. Using the social ecological model to understand AI use among youth potentially informs prevention of problematic outcomes and development of assets. Within a youth's ecology they experience individual interaction with their immediate environment, or *microsystem*, this includes peers, family, school, and workplaces and bring with them certain characteristics such as gender, age, and race or ethnicity. Research on youth AI use and characteristics linked to problems is in its infancy. However, studies are examining its effects on young children in experimental (Xu *et al.*, 2025) and school settings (Lee *et al.*, 2025). For the familial microsystem, recent research published in the journal *Family Relations*, examines potentially healthy forms of family use of AI with parents and children participating together in activities (McDaniel *et al.*, 2025; Wald *et al.*, 2025). Family monitoring and support of family AI use together with children may be a protective factor against problematic outcomes. Most relevant to community psychology is the role of the *exosystem* within the youth's ecology, with indirect yet significant influence on the child. This could include neighborhood environment, school climate, and parents' workplace. A parent who believes they lost a job due to AI or another who has a workplace supporting their positive interactions with AI to complete tasks at work, will differently influence how their children view and interact with AI. Beyond the exosystem is the *macrosystem* or sociocultural milieu, including culture and laws sanctioning AI, such as implementing policies that protect youth from problematic use. The chronosystem as a concept of the passage of time can include intraindividual as well as contextual change. When an adolescent moves into emerging adulthood and higher education or the workforce, societal developments in technology as well as expectations for appropriate AI use ensue. For example, adolescents in secondary school may need to learn more sophisticated, ethical ways to use AI in their schoolwork or careers from teachers, supervisors, or colleagues. Understanding the interactions among levels or systems to explain potentials of AI use on youth outcomes will guide policy and practice in families, institutions they attend and work at, and communities at large.

8.3 The promise of computational ethics

A Piagetian explanation of how AI learns to reason morally is not feasible for applications in machine learning to keeping youth safe. However, a contextual worldview, specifically the social ecological framework, was explored as a promising application to understanding and promoting youths' healthful and safe use of AI.

Further, in recent literature on developing moral AI, working within a computational ethics paradigm holds promise.

AI currently has limited capacity to take on ethical and moral reasoning challenges; hence Awad and colleagues (2022) recommended that morality be algorithmically determined. An example Awad *et al.* used to illustrate developing machine learning (ML) capacities for human capabilities, was how AI gained visual perception performance that exceeded that of human doctors for identifying cancer through scans. This innovation was determined by algorithm, where the human mind was essentially programmed into the machine (Awad *et al.*, 2022, p. 388).

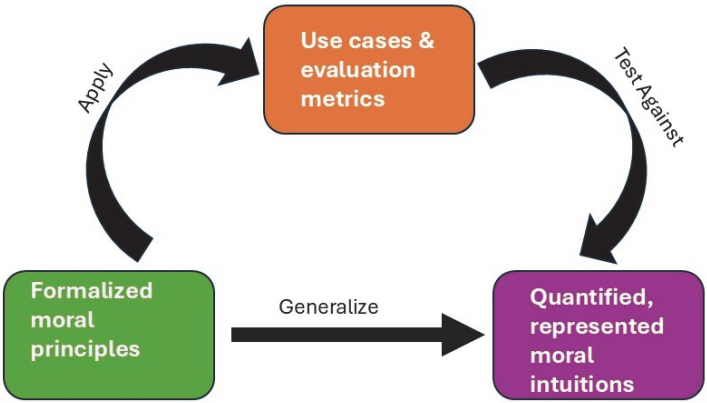
Machine ethics entails developing AI systems that behave, including the making of decisions, in ethically acceptable ways. A starting point is to apply algorithmic accountability to the process of developing an AI system wherein the purpose, structure, and behaviors of algorithms is transparent and efforts are made to mitigate bias (Awad *et al.*, 2022).

Their proposed computational ethics framework is illustrated by its most basic model of the reflective equilibrium framework wherein moral intuitions are generalized into moral principles. Moral principles are applied to specific cases and tested against moral intuitions for validity.

This framework is operative in moral philosophy and is a feedback loop taking the abstract principles to the more concrete cases (deductive process) and back again through an inductive process. Awad and colleagues propose a more complex version of Figure 1 for computational application.

Going through the more complex version of their model is beyond the scope of this paper. Rather, the promise their work poses for application in ML of moral understanding and behavior gives one hope for increasing the capacity for AI to develop a "conscience" and protect young users from harm while maximizing benefits.

Fig. 1 – Reflective equilibrium framework. This framework involves bringing moral principles and moral intuitions into alignment with one another through the use of examples or cases to which the moral principles are applied, and against which the moral intuitions are tested. (adapted from Awad et al.*, 2022)*



8.4 Morality as cooperation

In addition to a computational ethics paradigm, Bridge and colleagues (2021) introduce machine ethics in AI with the intention to create machines that are capable of moral reasoning and decision-making. A top-down or deductive approach to moral learning, by having a set of rules in place for machine learning, poses challenge on deciding what principles or values should be primary. For example, Kohlberg used an ethic of justice framework to evaluate levels of moral reasoning whereas Gilligan provided a feminist counterpoint with an ethic of care as the standard for morality (Thomas, 2005). The inductive or bottom-up approach for training AI is similarly problematic to translate learning into a general applicable heuristic or rule for future behavior. According to Oliver Bridge and his colleagues at Oxford (2021), training AI to develop morally through trust has great promise, similar to the process of moral socialization among youth. Bridge and colleagues (2021) main criteria or proposed standard is morality as cooperation, a cross-cultural conception of moral reasoning. This holds additional promise as Kohlberg and Gilligan were critiqued for lack of inclusion of specific adolescent populations in their research as Kohlberg's theory was based on research with males and Gilligan's alternative to Kohlberg's theory was developed studying Caucasian middle-class female adolescents. Neither included

youth from diverse racial, cultural, and ethnic backgrounds. Morality as cooperation contains 7 components or types of cooperation, including: (1) allocating resources to kin; (2) coordinating mutual benefit; (3) social exchange/reciprocation; (4) bravery; (5) respecting those in higher positions; (6) dividing resources; and (7) respecting prior possession. Bridge and colleagues developed a curriculum indicating the behaviors an AI chatbot must learn and model from humans, particularly community and familial leaders. This solution poses a unique approach beyond developmental psychology, and the work of Piaget, Kohlberg, and Gilligan, that extends into the field of community psychology for AI applications. Community psychology supports collaborative approaches including models of inquiry such as community-based participatory action research (CBPAR) or youth participatory action research (YPAR) as a specific form of CBPAR. These collaborative approaches reject the role of "knowledgeable experts" with hierarchical approaches of community-based intervention and are inclusive of multiple voices for addressing community concerns. And, specific to youth, YPAR combats premises of "adultism" wherein adults and youth collaborate and share power in making community change (Ozer *et al.*, 2020). Last, AI tools must be taught humility in the conveying of information in their possession. Treating AI as a knowledgeable higher entity, or worse where AI tells their human companion they are a god, is where human-chatbot interactions go astray.

8.5 Solution two: the promise of AI chatbots in educational settings

Although recent literature finds AI is not ready for full-on application in therapeutic settings (Moore *et al.*, 2025), there is a growing body of literature for positive outcomes gleaned by children who interact with social and instructional robots in formal educational settings. The commercial use and creation of chatbots such as CharacterAI and Replika, had seemingly well-intentioned developers who likely did not consult with experts and practitioners in youth development, mental health, nor engaged in youth participatory approaches for their creation. The bottom line with commercial AI use and related algorithms is capital gain (Fisher, 2022). The use of social robots in childhood education reveals over a decade of positive outcomes for users indicative of beneficence (Smakman *et al.*, 2021). Social robots are a unique type of AI chatbot that interact in human-like ways, displaying social skills with capacities to read emotions and engage in conversation. They serve as peers, tutors, or even mentors to children and provide numerous academic and social relational benefits. For example, AI humanoid robots used with

low-income 5-year-olds living in rural South Korea experienced increases in AI literacy, with both teachers and parents attributing gains in children's cognitive skills, such as problem solving and science knowledge, to their interactions with the robots (Lee *et al.*, 2025). Moreover, children who engage with social robots report greater enjoyment of learning, taking in new information more relevant to their style of processing than with traditional classroom instruction and in outside school settings. Teachers benefit from students' use of social robots with lesser administrative burden and increased job satisfaction. Robots with physical presence, versus a tablet-based or two-dimensional delivery, produce more enjoyment for students. Stakeholders such as parents, teachers, and school administrators have a lot of trust for robots. Concerns remain, particularly in European educational studies, calling for the need for ethical guidelines for the use and deployment of social robots with younger children (Smakman *et al.*, 2021).

9. Conclusions and current progress

The dangers of unsupervised, unregulated chatbot use for children, adolescents, and emerging adults are noteworthy. A pandemic-based cohort effect may be happening with adolescents and emerging adults with proneness to seek out AI companionship through chatbots, gaming, and fantasy-based applications. As humans are wired for attachment and assigning human characteristics to inanimate objects, the promise and pitfalls of AI are ever looming. At this time approaches to teaching AI tools to reason and behave morally are in their infancy, supported by theory to be used to guide training, such as the seven principles from Bridge and colleagues' (2021) proposed morality as cooperation and Awad and colleagues' (2022) computational ethics frameworks. Current recommendations include applying developmental theory from a contextual world view, such as Bronfenbrenner's social ecological model. Also, conceptualizing AI morality with multiple social contextual criteria within the field of community psychology is useful, as contrasted with individualistic psychological paradigms. More than a decade of evidence stands out for effective and ethical uses of AI robots with children in educational settings (Smakman *et al.*, 2021), that include parent and teacher supervision and evaluation of students' learning. Additional positive uses of AI with youth are highlighted in research on youth with disabilities and other special needs (Ramadan *et al.*, 2020; Zhou *et al.*, 2023) as well as for sexual and gender minority young adults to promote sexual health and well-being (Bragazzi *et al.*, 2023). Beyond clinical and educational applica-

tions of AI companions, commercial applications used by youth and emerging adults continue to receive negative media attention and research mostly supports its use in association with poor mental health consequences. There are minor exceptions such as findings on young adult users' endorsements of Replika for friendship and companionship (Ta *et al.*, 2020), with mixed results on self-reported perceptions of its benefits (Bowen & Watson, 2024; De Freitas *et al.*, 2024). And, last, there is promise in the theory of mind research from an information processing understanding of child development wherein young children can distinguish between AI and human interactions (Xu *et al.*, 2025).

10. Future directions

Healthy interpersonal interactions are foundational for youth development. This could translate into balanced online or virtual interactions with interpersonal social communications involving other humans. Healthy relationship education for youth taking place in-person that includes discussions of partnership and marriage, has promoted positive lasting effects into adulthood (Administration for Children and Families, 2024). Relationship education for youth should consider introducing curriculum content on the pitfalls of romantic or other significant relationships with AI chatbots, providing ways to establish healthy boundaries and interactions with AI, that can be maintained through shared human-robot spaces with parents, peers, and mental health professionals. For example, there is evidence of healthful familial interactions with AI (McDaniel *et al.*, 2025; Wald *et al.*, 2025). Next steps to ensure children's safety with AI use are to challenge software developers to code for a morally socialized AI within LLM (Awad *et al.*, 2023), including youth and teacher input within community-based networks. Humans and AIs alike, with multiple AI tools communicating with and supervising each other, could serve to monitor AI behaviors as moral decisions are being made. With current restrictions in place to ensure safe social robot use for younger children (Lee *et al.*, 2025), exemplars can be applied to promoting safe AI use for adolescents, especially promoting its ethical use academically and for promoting career readiness. Policy is vital as the United States lacks regulatory capacity on AI and children's use (policies vary by state) prompted by capitalist priorities (surveillance and commercialism – see Zuboff, 2019 and Fisher, 2022). Italy and other European countries place age restrictions, even bans, on AI applications with potential to cause harm. Last, partnerships, or at least consults, between developmental and community psychologists as

advisors to AI program developers, combined with use of appropriate computational ethics paradigms (e.g., Awad *et al.*, 2023) to make purposes and algorithms transparent within applications, are recommended future directions to produce ethical AI in promoting positive youth development. Networks of concerned parents, adults, and young people can work together, through assistance of community organizers and via participatory approaches to implement change through community-based education and policy enactment.

References

Administration for Children & Families, Office of Family Assistance, & Chamberlain, S. (2024). *Promoting Healthy Youth Relationships*. Webinar Attended December 12, 2024.

Asimov, I. (1942). Runaround. *Astounding Science Fiction, 29*(1), 94-103.

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A. C., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., Mikhail, J., Opoku-Agyemang, K., Borg, J. S., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., & Tenenbaum, J. B. (2022). Computational ethics. *Trends in Cognitive Sciences, 26*(5), 388-405. https://doi.org/10.1016/j.tics.2022.02.009.

Bowen, J. A. & Watson, C. E. (2024). *Teaching with AI: A practical guide to a new era of human learning.* Johns Hopkins University.

Bragazzi, N. L., Crapanzano, A., Converti, M., Zerbetto, R., & Khamisy-Farah, R. (2023). The impact of generative conversational AI on the lesbian, gay, bisexual, transgender, and queer community: Scoping review. *Journal of Medical Internet Research, 25,* Article e52091. https://doi.org/10.2196/52091.

Bridge, O., Raper, R., Strong, N., & Nugent, S. E. (2021). Modelling a socialised chatbot using trust development in children: Lessons learnt from Tay. *Cognitive Computation and Systems, 3*, 101-108. https://doi.org/10.1049/ccs2.12019.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, January-June*, 1-12. https://doi.org/10.1177/2053951715622512.

Chow, A. R. (2023, February 23). AI–human romances are flourishing and this is just the beginning. *Time*. https://time.com/6257790/ai-chatbots-love

Crain, W. (2014). *Theories of Development: Concepts and Applications.* (6th Ed.) Prentice Hall.

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

De Freitas, J., Castelo, N., Uğuralp, A. K., & Uğuralp, Z. (2024). Lessons from an app update at Replika AI: Identity discontinuity in human–AI relationships. *Harvard Business Working Paper No. 25-018*. https://doi.org/10.2139/ssrn.4976449.

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

Espelage, D. L. (2014). Ecological theory: Preventing youth bullying, aggression, and victimization. *Theory Into Practice, 53*, 257-264. https://doi.org/10.1080/00405841.2014.947216.

Fisher, M. (2022). The chaos machine: *The inside story of how social media rewired our minds and our world*. Little, Brown and Company.

Guerreiro, J., & Loureiro, S. M. C. (2023). I am attracted to my cool smart assistant! Analyzing attachment-aversion in AI–human relationships. *Journal of Business Research, 161*, Article 113863. https://doi.org/10.1016/j.jbusres.2023.113863.

Kumar, S. & Choudhury, S. (2023). Cognitive morality and Artificial Intelligence (AI): A proposed classification of AI systems using Kohlberg's theory of cognitive ethics. *Technological Sustainability*. http://dx.doi.org/10.2139/ssrn.4293968

Leaver, T. (2015). Born Digital? Presence, Privacy, and Intimate Surveillance. In Hartley, J. & W. Qu (Eds.), *Re-Orientation: Translingual Transcultural Transmedia. Studies in narrative, language, identity, and knowledge* (pp. 149–160). Fudan University Press.

Lee, B., Ku, S., & Ko, K. (2025). AI robots promote South Korean preschoolers' AI literacy and computational thinking. *Family Relations, 74*(3), 1354-1375. https://doi.org/10.1111/fare.13189.

Lee, H. Y., Kim, I., & Kim, J. (2024). Adolescents' Mental Health Concerns in Pre- and During COVID-19: Roles of Adverse Childhood Experiences and Emotional Resilience. *Child Psychiatry & Human Development,* 1-12. https://doi.org/10.1007/s10578-024-01726-x.

Liu, Y., Mittal, A., Yang, D., & Bruckman, A. (2022). Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. *CHI Conference on Human Factors in Computing Systems,* https://doi.org/10.1145/3491102.3517731.

Mansoor, M., Hamide, A., & Tran, T. (2025). Conversational AI in pediatric mental health: A narrative review. *Children, 12*, 359-381. https://doi.org/10.3390/children12030359

Martens, M., Abeel, M. V., & De Wolf, R. (2025). Home maintainer, guardian or companion? Three commentaries on the implications of domestic AI in the household. *Family Relations, 74*(3), 1098-1108. https://doi.org/10.1111/fare.13162

McDaniel, B. T., Coupe, A., Weston, A., & Pater, J. A. (2025). Emerging Ideas. A brief commentary on human–AI attachment and possible impacts on family dynamics. *Family Relations, 74*(3), 1072–1079. https://doi.org/10.1111/fare.13188.

Montreuil, M., Gendron-Cloutier, L., Laberge-Perrault, E., Piché, G., Genest, C., Rassy, J., Malboeuf-Hurtubise, C., Gilbert, E., Bogossian, A., Camden, C., Mastine, T., & Barbo, C. (2023). Children's and adolescents' mental health during the COVID-19 pandemic: A qualitative study of their experiences. *Journal of Child & Adolescent Psychiatric Nursing, 36*(2), 65-74. https://doi.org/10.1111/jcap.12404.

Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (April, 2025). *Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers*. International Conference on Learning Representations. Singapore.

National Institute on Drug Abuse (2023). *Covid-19 and Substance Use*. Accessed from https://nida.nih.gov/research-topics/covid-19-substance-use.

Oh, J., Kim, M., Rhee, S. Y., Rahmati, M., Koyanagi, A. Smith, L., Kim, M. S., Fond, G., Boyer, L., Kim, S., Shin, J. I., & Yon, D. K. (2024). National trends in the prevalence of screen time and its association with biopsychosocial risk factors among Korean adolescents, 2008–2021. *Journal of Adolescent Health*, *74*(3), 504-513. https://doi.org/10.1016/j.jadohealth.2023.10.021.

Ozer, E., Abraczinskas, M., Duarte, C., Mathur, R., Ballard, P. J., Gibbs, L., Olivas, E. T., Bewa, M. J., & Afifi, R, (2020). Youth participatory approaches and health equity: Conceptualization and integrative review. *American Journal of Community Psychology, 66*, 267-278. https://doi.org/10.1002/ajcp.12451.

Pepper, S. C. (1942). *World hypotheses: A study in evidence*. University of California Press.

Ramadan, Z., Farah, M. F., & El Essrawi, L. (2020). From Amazon.com to Amazon.love: How Alexa is redefining companionship and interdependence for people with special needs. *Psychology of Marketing, 38*, 596-609. https://doi.org/10.1002/mar.21441

Smakman, M., Vogt, P., & Konijn, E. A. (2021). Moral considerations on social robots in education: A multi-stakeholder perspective. *Computers and Education, 174*, 1-14. https://doi.org/10.1016/j.compedu.2021.104317.

Song, X., Xu, B., & Zhao, Z. (2022). Can people experience romantic love for AI? An empirical study of intelligent assistants. *Information & Management, 59,* (2022), 1-10. https://doi.org/10.1016/j.im.2022.103595.

Strasburger, V. C., & Wilson, B. J. (2002). *Children, Adolescents & the Media*. Sage Publications.

Sullivan, H. S. (1953). *The interpersonal theory of Psychiatry*. Routledge.

Thomas, R. M. (2005). *Comparing Theories of Child Development*: Wadsworth.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research, 22*(3), Article e16235, 1-10. https://doi.org/10.2196/16235.

Treiman, L., Ho, C. J., & Kool, W. (2024). The consequences of AI training on human decision-making. *Psychological & Cognitive Sciences, 121*(33), 1-12. https://doi.org/10.1073/pnas.2408731121.

Tudge, J. R. H., Payir, A., Mercon-Vargas, E., Cao, H., Liang, Y., Li, J., & O'Brien, L. (2016). Still misused after all these years? A reevaluation of the uses of Bronfenbrenner's bioecological theory of human development. *Journal of Family Theory & Review, 8,* (December 2016), 427-445. https://doi.org/10.1111/jftr.12165.

Turkle, S. (2024, March 27). Who do we become when we talk to machines? An MIT exploration of generative AI. 10.21428/e4baedd9.caa10d84

Woodside, M. R. & McClam, T. (2019). *An Introduction to Human Services*, *9th Edition*. Cengage Publishing.

Valz, D. (2023). Personalization: *Why the relational modes between generative AI Chatbots and human users are critical factors for product design and safety* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.4468899.

Wald, R., Piotrowski, J. T., Araujo, T., & van Oosten, J. M. F. (2023). Virtual assistants in the family home. Understanding parents' motivations to use virtual assistants with their children. *Computers in Human Behavior, 139*, 1-12. https://doi.org/10.1016/j.chb.2022.107526.

Williams, M., Carroll, M., Narang, A., Weisser, C., Murphy, B., & Dragan, A. (2025). On targeted manipulation and deception when optimizing LLMs for user feedback. *International Conference on Learning Representations*. Singapore.

Xu, Y., Thomas, T., Yu, C. L., & Pan, E. Z. (2025). What makes children perceive or not perceive minds in generative AI? *Computers in Human Behavior: Artificial Humans, 4*(2025), 100135. https://doi.org/10.1016/j.chbah.2025.100135.

Zhang, S., Li, J., Cagiltay, B., Kirkorian, H., Mutlu, B., & Fawaz, K. (2025). A qualitative exploration of parents and their children's uses and gratifications of ChatGPT. *Family Relations, 74*, 1056-1071. https://doi.org/10.1111/fare.13171.

Zhou, H., Wu, X., & Yu, L. (2023). The comforting companion: Using AI to bring loved one's voices to newborns, infants, and unconscious patients in ICU. *Critical Care, 27*(1), Article 135. https://doi.org/10.1186/s13054-023-04418-5.

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