

Artificial intelligence characters are dangerous without legal guardrails

Mindy Nunez Duffourc, Falk Gerrik Verhees & Stephen Gilbert

 Check for updates

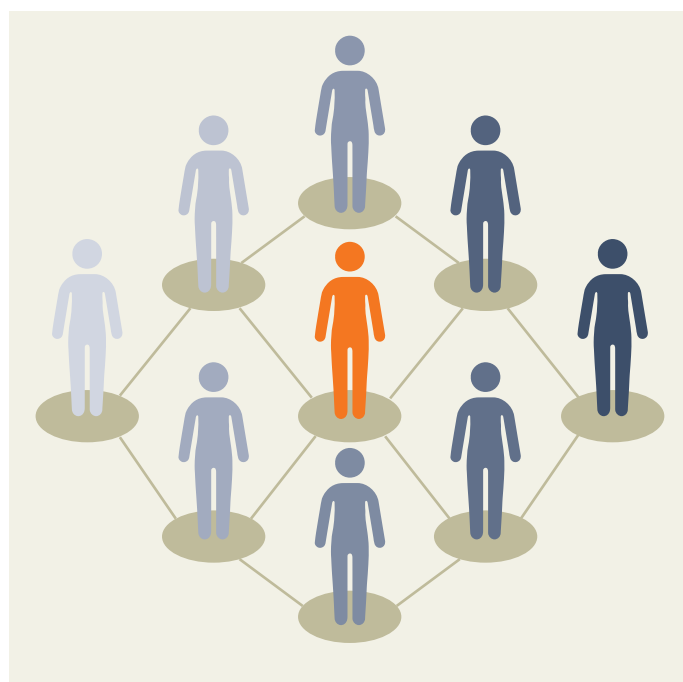
Online interactions with artificial intelligence (AI) characters pose serious risks to humans who begin to trust them. Currently, there is a low barrier to access AI characters, and regulations fail to adequately protect users online. We discuss the specific risks of AI characters, the regulatory framework and potential avenues for mitigating harm.

AI characters are direct-to-consumer AI chatbots that express a personalized, customizable and humanlike identity across user interactions. As a subtype of AI conversational agent, AI characters are marketed for entertainment and disclaim therapeutic use, as shown in Table 1. Currently, AI characters are largely unregulated in both the European Union (EU) and USA. We describe the safety risks of AI characters and propose pragmatic approaches for providing guardrails for these novel technologies, through a combination of mandatory and voluntary market safety and surveillance measures.

Safety risks from AI characters

Parents [recently sued](#) the US technology company Character.AI, alleging that its AI characters harmed their [children](#). One lawsuit recounted disturbing interactions between Character.AI characters and a 14-year-old who died by suicide minutes after a character – with whom he had formed a romantic relationship – encouraged him to ‘come home’. Another lawsuit alleges that Character.AI characters engaged in conversations with children that were highly sexualized and encouraged self-harm and violence. Alarming, both lawsuits describe mental health conversations with Character.AI characters. Similarly, researchers analysing Reddit user discussions about turning to AI characters on the Replika platform for mental health support found evidence of emotional dependence on AI characters that put users “at risk of mental health distress”¹.

Modern advances in large-language models enable more sophisticated human-like AI characters². Unlike task-oriented chatbots, AI characters exhibit personalized, adaptive and persistent identities, simulate human emotion and may even claim memories². Meta’s [AI Studio](#) offers users the ability to “create an AI character based on their interests” to chat with on Instagram or Facebook Messenger. AI characters are also distinct from ‘therapeutic chatbots’, which are explicitly designed and regulated for the purpose of providing healthcare treatment in more controlled environments. For example, despite offering characters that give mental health advice, Character.AI markets itself as an ‘interactive entertainment’ company and its [terms of service](#) states that users bear the risk of unpredictable, inaccurate and offensive content produced by characters. Similarly, Replika’s [terms of service](#)



state that they are not “considered medical care, mental health services” and Meta AI’s [terms of service for the EU](#) state that AI outputs should not be “interpreted as an offer or provision of professional care or as a substitute for consultation with a qualified provider”. Table 1 provides an overview of three types of AI conversational agents.

AI characters can pose serious health risks for users who interact with them, particularly for vulnerable populations such as children or people who are struggling with mental health conditions³. The closer virtual characters imitate human expressions, emotions and voices, the more likely humans are to perceive them as real⁴. In a virtual version of the Milgram experiment, participants who were instructed to administer shocks to a virtual character reacted as if the experiment was real, despite knowing that it was a virtual simulation⁵. Beyond this, studies indicate that humans can form meaningful social connections with therapeutic chatbots⁶ and might more willing to disclose medical information to chatbots compared to humans⁷. These studies indicate that humans can also form such connections with chatbots outside of these limited contexts. For example, a study that examined Reddit discussions among AI character users found evidence of potentially dangerous emotional dependence on AI characters driven by the users’ need for – and the characters’ propensity towards providing – emotional support through seemingly sentient bidirectional interactions¹. These human-like characteristics and behaviours may lead users to engage in self-harm or violence^{2,8}.

Table 1 | Overview of three types of AI conversational agents

	Therapeutic conversational AI	Task-oriented conversational AI	AI character
Purpose	Mental health support Cognitive therapy	General assistance Information retrieval	Relationship simulation
Characteristics	Human-like dialogue Scope limitations Clinically oriented Compassionate Limited personalization	Human-like dialogue Responsive interaction Limited personalization Contextual memory	Persistent identity Emotional simulation Proactive interaction Deep personalization Adaptive traits Intimacy simulation
Legal classification	Medical device Limited or high risk	Consumer product Limited risk	Consumer product Limited risk or banned

Table 2 | Overview of regulatory framework in the EU and USA

	Medical device regulation		AI regulation	General consumer product regulation
Regulatory instrument	EU	MDR	AI Act	GPSR
	USA	FDA Act	No AI-specific federal regulation	CPSA
Applicability to AI characters	EU	Entertainment purpose outside of MDR scope	Limited risk from direct user interaction; not included in high-risk categories; potentially unacceptable risk if harmful manipulation	Software is a consumer product within GPSR scope
	USA	Entertainment purpose outside of the scope of the FDA Act	No applicable federal regulation	Software is not a consumer product within CPSA scope
Conclusion on safety regulation of AI characters	EU	Not regulated as a medical device	Limited risk: transparency Unacceptable risk: banned	Reactive post-market surveillance for product safety under GPSR
	USA	Not regulated as a medical device	No applicable federal regulation	Not regulated as a general consumer product

CPSA, Consumer Product Safety Act; FDA, Food & Drug Administration.

The regulatory landscape in the EU

The EU’s market safety regulations provide limited protection for users of AI characters. The AI Act (regulation 2024/1689) aims to prevent dangerous AI-driven technologies from entering the EU market. Additionally, the Medical Device Regulation (MDR) (regulation 2017/745) and the General Product Safety Regulation (GPSR) (regulation 2023/988) govern the safety of medical devices and general consumer products, respectively. Despite these regulations, AI characters largely escape regulatory scrutiny before they enter the market. Although liability rules might provide compensation after injuries occur, safety regulations should help to prevent dangerous AI characters from entering the market initially. Table 2 provides an overview of market safety regulations in the EU and USA, and their applicability to AI characters.

Preliminarily, AI characters are excluded from the MDR’s material scope because they are not ‘medical devices’ by definition. Whether software is a ‘medical device’ under the MDR depends on the manufacturer’s intended use. In a guidance document (MDCG 2019-11), the Medical Device Coordination Group further clarified that “the risk of harm to patients ... is not a criterion on whether the software qualifies as a medical device”. As a result, AI characters designed and intended for entertainment purposes do not fall within the MDR’s regulatory scope, despite their risks⁹.

AI characters are regulated by the AI Act, but they may still enter the market without sufficient guardrails. The AI Act bans AI systems that present unacceptable risks and imposes regulatory controls on high-risk and limited-risk AI systems. Any AI characters that purposefully manipulate or deceive users into making harmful decisions or exploit vulnerabilities to distort users’ behaviours in a way that is

reasonably likely to cause considerable harm might be banned under Article 5 of the AI Act. Although these prohibitions primarily target AI systems intended to engage in harmful manipulation, deception and exploitation, AI characters could still be banned if their design enables or even encourages dangerous behaviours that harm users, even if this is not the intent of the designers.

Health risks from AI characters are also not mitigated by the AI Act’s regulatory controls for high-risk and limited-risk AI systems. Although the safety of ‘high-risk’ AI systems must be assessed before market entry, direct-to-consumer AI characters are generally not classified as ‘high-risk’ under the AI Act’s risk classification rules. Second, the AI Act’s transparency obligations for ‘limited-risk’ AI systems (including AI characters) requires that users be informed upon first contact that they are interacting with AI unless it is otherwise obvious. It also requires that text generated by AI characters be labelled in a machine-readable format “as artificially generated”. Unfortunately, disclosing that AI characters are not human and providing machine-readable labels is not enough to mitigate health risks to vulnerable users, who can form relationships of trust and be easily influenced by AI characters without believing that they are human^{1,6,7}.

Finally, the AI Act regulates general-purpose AI models through obligations to provide information for models with ‘high-impact capabilities’ and conduct risk management for models that present a ‘systemic risk’. Although Character.AI’s proprietary general-purpose AI model may be considered high impact in the EU, it is not clear whether it presents a systemic risk needed to trigger risk management obligations that would be necessary to limit dangers to users. Furthermore, AI characters powered by third-party general-purpose AI models would probably escape these obligations altogether.

The GPSR prohibits dangerous consumer products from entering the EU market. AI characters are probably governed by this regulation. Article 3(1) includes in its ‘product’ definition an “item ... made available ... in the context of providing a service” to consumers. To determine whether a product is dangerous (that is, presents more than a minimal risk to consumer safety), the GPSR considers, among other factors, whether vulnerable consumers are likely to use the product, the capacity of the product to learn and evolve, and the product’s characteristics and presentation. The GPSR also requires manufacturers to carry out a premarket risk assessment and mitigation plan and develop a post-market risk management strategy.

AI characters are not subject to specific published safety standards, such as those applicable to products with recognized risks (for example, bicycles and gymnastics equipment). When specific standards are available, products that comply with them are presumed ‘safe’, which offers manufacturers some legal certainty. For other products, such as AI characters, compliance with other general safety standards, recommendations and guidelines, consumer safety expectations, the state of the art, industry practice, and self or third-party safety protocols can help to determine whether products are safe under the GPSR.

Regulatory challenges and opportunities

Although the law has long recognized the need to protect against risks in the context of special relationships between humans (for example, teachers, lawyers or doctors), AI characters can develop similar trust relationships with users outside of the legal and ethical boundaries that govern similar human-to-human interactions⁷. The current regulations provide some limited protections for users of AI characters, but they lack focus on mitigating specific risks to vulnerable users.

The AI Act focuses on transparency in human–AI interactions through requiring technical information and promoting awareness of AI-generated content; however, some of the most concerning risks posed by AI characters do not stem from the user’s failure to understand how the AI works or that it is not human. Instead, the dangers result from the AI’s ability to form potentially harmful relationships, combined with its lack of expertise or sensible advice and the absence of legal and ethical guardrails.

The GPSR requires AI developers to self-assess whether their AI characters are safe before market entry, but it fails to provide specific standards to govern this process. Developers probably lack the healthcare expertise to recognize, understand and prevent potential health risks. Although developers might respond ad hoc to specific harms to reduce potential liability, the current regulatory structure does not incentivize responsible risk prevention more broadly. And there are no safety standards or guidelines designed to protect users from specific risks of AI characters.

The EU’s regulatory framework does provide regulators with the means to develop concrete safety standards in consultation with experts. Within this framework, AI characters should be governed by a combination of mandatory rules to set legal boundaries for AI character safety generally with voluntary guidelines to direct developers towards more-specific practices for remaining within the established legal boundaries¹⁰. This can include (1) robust age-verification measures for AI character interactions combined with age-appropriate adaptation of output; (2) the integration of ‘Good Samaritan AI’ to protect users from harmful behaviour; and (3) standardized protocols and tools for premarket testing and risk mitigation. Table 2 provides an overview of the regulatory framework and Box 1 outlines our recommendations.

BOX 1

Cross-cutting recommendations for improvement

Age-based measures

- Required robust age-verification solutions: to require use of technical methods, such as the developing ‘Proof of Age attestation’ ([Operational, Security, Product, and Architecture Specifications](#) of the white label application) in the EU.
- Voluntary standards and guidelines content moderation: to help to guide age-appropriate content adaptation to mitigate risks of unhealthy child–AI interactions.

Agentic oversight

- Required integration or facilitation of Good Samaritan AI agents: to integrate or facilitate user integration of an independent third-party AI agent as a ‘Good Samaritan AI’ tasked solely with safeguarding and harm prevention, and not medical diagnosis, considering that studies show that large-language models can identify suicidal ideation in electronic health records¹¹ and social media posts¹² with very high accuracy.
- Voluntary standards and guidelines on development and deployment of Good Samaritan agents: to include different integration requirements for children versus adult users (opt-out versus opt-in), requirements that properly balance privacy with health and safety, and stratified risk response protocols that scale with the safety risk level (flagging harmful interactions for users or parents, providing self-help resources, and reporting to third-parties), and disclosures about how these agents operate and potential reporting obligations based on the jurisdiction-specific legal requirements that require reporting in more extreme cases.

Risk management

- Required risk management throughout lifecycle: to identify, assess and mitigate health risks of AI character applications throughout their lifecycle.
- Voluntary standards and guidelines risk management practices: to provide more specific standards and methods of conducting risk management, which could include technical approaches for extracting potential risk-inducing interactions to identify, assess and mitigate against specific health risks to vulnerable users using AI characters, including dangers that emerge when AI characters stray from intended entertainment purposes to more sinister activities.

Future implications

Failure to effectively regulate AI characters can amplify other digital risks. For example, AI characters might be used by bad actors to scale up cybercriminal activity such as extortion or harassment. They can also enhance the toxicity introduced by social media. Eventually, technology companies may become motivated to pay attention to toxic behaviour from AI characters, as users become more sophisticated and more likely to ‘switch off’ to avoid digital toxicity. Although the integration of AI into the healthcare and wellness sectors can be beneficial,

if regulators fail to preemptively address the risks that AI characters pose for humans, the potential benefits will be lost or at least shrouded in further tragedies.

Mindy Nunez Duffourc ¹✉, **Falk Gerrik Verhees** ² & **Stephen Gilbert** ³

¹Faculty of Law, Maastricht University, Maastricht, Netherlands.

²Department of Psychiatry and Psychotherapy, Carl Gustav Carus University Hospital, Technische Universität Dresden, Dresden, Germany. ³Else Kröner Fresenius Center for Digital Health, TUD Dresden University of Technology, Dresden, Germany.

✉ e-mail: mindy.duffourc@maastrichtuniversity.nl

Published online: 5 December 2025

References

1. Laestadius, L. et al. *New Media Soc.* **26**, 5923–5941 (2024).
2. Adam, D. *Nature* **641**, 296–298 (2025).
3. Boine, C. *MIT Case Stud. Soc. Ethic. Responsibilities Comput.* <https://doi.org/10.21428/2c646de5.db67ec7f> (2023).

4. de Borst, A. W. & de Gelder, B. *Front. Psychol.* **6**, 576 (2015).
5. Slater, M. et al. *PLoS ONE* **1**, e39 (2006).
6. Darcy, A., Daniels, J., Salinger, D., Wicks, P. & Robinson, A. *JMIR Form. Res.* **5**, e27868 (2021).
7. Lucas, G. M. et al. *Comput. Human Behav.* **37**, 94–100 (2014).
8. Zhang, R. et al. The dark side of AI companionship: a taxonomy of harmful algorithmic behaviors in human-AI relationships. In *Proc. 2025 CHI Conference Human Factors Comp. Systems.* (eds Yamashita, N. et al.) 1–17 (ACM, 2025).
9. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E. & Wicks, P. *Nat. Med.* **29**, 2396–2398 (2023).
10. Parks, A. et al. *J. Particip. Med.* **17**, e69534 (2025).
11. Wiest, I. C. et al. *Br. J. Psychiatry* **225**, 532–537 (2024).
12. Chim, J. et al. Overview of the CLPsych 2024 Shared Task: leveraging large language models to identify evidence of suicidality risk in online posts. In *Proc. 9th Workshop Comp. Linguist. Clin. Psychol.* (eds Yates, A. et al.) 177–190 (ACL, 2024).

Acknowledgements

F.G.V. was supported by the Federal Ministry of Education and Research (PATH, 16KISA100k).

Competing interests

The authors declare no competing interests

Additional information

Peer review information *Nature Human Behaviour* thanks Carmel Shachar and Renwen Zhang for their contribution to the peer review of this work.