frontiers | Research Topics

Ethical design of artificial intelligence-based systems for decision making

Edited by
Valentina Franzoni, Jordi Vallverdu, Roberto Capobianco, Giulio Biondi, Alfredo Milani, Francesca Alessandra Lisi and Stefano Cagnoni

Published in
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-1714-7
DOI 10.3389/978-2-8325-1714-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Ethical design of artificial intelligence-based systems for decision making

Topic editors

Valentina Franzoni — University of Perugia, Italy
Jordi Vallverdu — Autonomous University of Barcelona, Spain
Roberto Capobianco — Sony AI, Switzerland
Giulio Biondi — University of Perugia, Italy
Alfredo Milani — University of Perugia, Italy
Francesca Alessandra Lisi — University of Bari Aldo Moro, Italy
Stefano Cagnoni — University of Parma, Italy

Citation

Franzoni, V., Vallverdu, J., Capobianco, R., Biondi, G., Milani, A., Lisi, F. A., Cagnoni, S., eds. (2023). *Ethical design of artificial intelligence-based systems for decision making*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-1714-7

Table of contents

frontiers | Frontiers in Artificial Intelligence

TYPE Editorial
PUBLISHED 24 July 2023
DOI 10.3389/frai.2023.1250209

OPEN ACCESS

EDITED AND REVIEWED BY
Julita Vassileva,
University of Saskatchewan, Canada

*CORRESPONDENCE
Valentina Franzoni
✉ valentina.franzoni@unipg.it

RECEIVED 29 June 2023
ACCEPTED 06 July 2023
PUBLISHED 24 July 2023

CITATION
Biondi G, Cagnoni S, Capobianco R, Franzoni V, Lisi FA, Milani A and Vallverdú J (2023) Editorial: Ethical design of artificial intelligence-based systems for decision making.
Front. Artif. Intell. 6:1250209.
doi: 10.3389/frai.2023.1250209

Editorial: Ethical design of artificial intelligence-based systems for decision making

Giulio Biondi[1], Stefano Cagnoni[2], Roberto Capobianco[3], Valentina Franzoni[1,4]*, Francesca A. Lisi[5], Alfredo Milani[1] and Jordi Vallverdú[6]

[1]EmoRe Research Group, Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy, [2]Department of Engineering and Architecture, University of Parma, Parma, Italy, [3]Artificial Intelligence and Robotics Research Group, Department of Computer, Control and Management Engineering, La Sapienza University of Rome, Rome, Italy, [4]Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China, [5]Department of Computer Science, University of Bari "Aldo Moro", Bari, Italy, [6]ICREA Acadèmia, Department of Philosophy, Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain

KEYWORDS

ethics, artificial intelligence, decision support, AI regulation, ethics in AI, fairness

Editorial on the Research Topic
Ethical design of artificial intelligence-based systems for decision making

Introduction

Emphasizing the importance of ethical design in AI-based decision-making systems is not only crucial from an emotional and social perspective but also from a legal and risk management standpoint (see Crawford and Calo, 2016). While EU regulations, such as Madiega (2021) or European Commission (2019), impact all Artificial Intelligence (AI) products in European countries, it is important to note that in the United States, AI regulations are voluntary and locally applied. On January 26, 2023, the National Institute of Standards and Technology (NIST), an agency of the US Department of Commerce, released the Artificial Intelligence Risk Management Framework 1.0 (RMF) (see Tabassi, 2023). This framework serves as a voluntary, non-sector-specific guide for technology companies engaged in the design, development, deployment, or utilization of AI systems. Its objective is to assist these companies in effectively managing the diverse risks associated with AI. AI technologies are subject to various legal frameworks and regulations that govern their use and mitigate potential risks. Ethical design ensures that AI systems comply with legal requirements, such as data privacy and protection laws, but also with human psychological and emotional needs (Vallverdú and Casacuberta, 2014; Franzoni and Milani, 2019); it incorporates mechanisms to safeguard personal information and ensure that AI systems operate within the bounds of legal frameworks (Coeckelbergh, 2020). Furthermore, ethical design considers risk management in the development and deployment of AI systems. It involves identifying and assessing potential risks associated with biases, discrimination, or unintended consequences (Buolamwini and Gebru, 2018; Biondi et al, 2022). By integrating risk management practices, such as rigorous testing, validation, and ongoing monitoring, the ethical design minimizes the likelihood of negative outcomes and helps mitigate legal liabilities, in both local and global domains (Jobin et al., 2019). On June 20th, 2023, the European Parliament made

significant progress in shaping the AI Act by adopting its negotiating position. This move aims to ensure that AI systems developed and utilized in Europe adhere to the principles and values of the European Union (EU), including human oversight, safety, privacy, transparency, non-discrimination, and social and environmental wellbeing. The Parliament's position highlights several key aspects. Firstly, they advocate for a complete ban on the use of AI for biometric surveillance, emotion recognition, and predictive policing. Secondly, they propose that generative AI systems, such as ChatGPT, should clearly disclose when content is AI-generated. Lastly, the Parliament considers AI systems used for influencing voters in elections as high-risk. The ethical design can also align with ethical guidelines and principles set forth by professional and regulatory bodies. Adhering to these guidelines promotes responsible and accountable use of AI technologies, reducing legal risks and ensuring compliance with industry standards. In summary, ethical design in AI-based decision-making systems goes hand in hand with legal compliance and risk management. It ensures that AI systems are developed and operated within legal boundaries, while also minimizing risks and liabilities. By embracing ethical principles, organizations can navigate the complex legal landscape surrounding AI technologies and mitigate potential legal and reputational risks associated with their deployment (see Vinuesa et al., 2020; Franzoni, 2023). By incorporating ethical considerations, AI-based decision-making systems can avoid perpetuating biases, discrimination, and other negative social consequences (see Biondi et al., 2022). Ethical design takes into account the diverse needs, preferences, and emotions of individuals, promoting inclusivity and fairness (Zafar et al., 2017). It recognizes the importance of transparency and interpretability, enabling users to understand and trust the decisions made by AI systems. Moreover, ethical design acknowledges the potential impact of AI decisions on social dynamics and relationships. It encourages responsible behavior and accountability, ensuring that AI systems are designed to align with societal norms and values. By prioritizing ethical design, we can ensure that AI technologies contribute positively to society while respecting the emotional and social fabric of human existence.

State of the art

Ethical design in AI-based decision-making systems is of paramount importance. Current approaches, methodologies, and frameworks address the ethical implications associated with these technologies. There are some fundamentals to be taken into account: Integration of fairness and non-discrimination principles promotes equitable outcomes and mitigates bias (Floridi et al., 2020); transparency and interpretability enhance trust and accountability (Larsson and Heintz, 2020); accountability ensures clear responsibility and mechanisms for addressing potential harms (Mittelstadt, 2019); privacy preservation techniques safeguard sensitive data while enabling collaboration (Manheim and Kaplan, 2019); and, finally, the ethical design fosters trust in AI technologies and mitigates unintended consequences (Bryson and Winfield, 2017). Challenges include balancing fairness and accuracy and addressing interpretability-performance trade-offs. Of course, practical and scalable frameworks are needed. Emphasizing ethical design in AI-based decision-making systems addresses societal concerns, reduces biases, enhances transparency, and establishes accountability (Novelli et al., 2023). Ongoing analysis promotes responsible AI systems aligned with societal values, benefiting individuals and communities. Therefore, exploring current approaches, methodologies, and frameworks in ethical design for AI systems is essential in addressing the ethical challenges posed by AI technologies. Researchers and practitioners have made significant strides in developing strategies to ensure responsible and accountable AI systems.

Research Topic on ethical design of artificial intelligence-based systems for decision making

Systematic reviews

In virtual educational settings, the impact of learner and teacher gender on human-to-human interaction and the persistence of gender stereotypes are of critical interest. In the systematic review of studies on Pedagogical Agents by Armando et al., authors discuss the impact of gender on learners' perception, academic performance, and self-evaluation skills. Findings indicate that male and female agents can improve performance, with female agents efficiently employable to contrast the stereotype threat, e.g., in male-dominated STEM environments. On the other hand, the agents' gender evidently impacts their pedagogical roles, appearance, and interaction patterns. Virtual agents whose gender does not match social stereotypes on context and roles may be less effective in conveying their messages e.g., older and elegant agents are perceived as experts; female agents are more successful in establishing positive relationships with learners. Androgynous systems as a potential solution require further investigation, as they may hinder efforts to avoid gender stereotypes. The review emphasizes the importance of gender choice and the need for further research in this area.

In the field of green economy and, in particular, regarding waste management applications, the review by Nkwo et al. highlights the significance of thoughtful and human-centered design in developing applications that raise awareness of social issues, using the Persuasive System Design (PSD) framework. The study investigates the incorporation and implementation of behavior change strategies and evaluates their effectiveness based on user ratings. The findings reveal that task-assistance strategies are prevalent, while credibility strategies enhance persuasiveness and trust. The impact of dialogue support strategies, feedback and reminder provisions, and social support strategies leveraging social influence across various dimensions, including app focus and waste management activities, correlate with app ratings. Based on the findings, the authors provide design suggestions and guidelines leveraging social influence e.g., sustainable waste management apps, emphasizing user-friendly routines, adaptive features, automated intelligent notifications, and performance tracking.

Novel research contributions

The three original research papers in this Research Topic (i.e., Thomas et al.; Chen et al.; Wang et al.) present contrasting viewpoints on user experience with digital interactive systems. Two papers analyze user behavior, while the third examines the impact of messages conveyed through such systems.

In Thomas et al., the authors critically review existing approaches to assessing message persuasiveness in different domains. As a result of their analysis, the authors propose and validate a new scale of persuasiveness based on user ratings of items from two domains: healthy eating advice and email security messages.

The other two papers focus on monitoring literacy learners' attention status and users' attitudes toward medical Artificial Intelligence. In Chen et al., the authors introduce a method to assess disengagement among literacy learners during online classes by measuring performance discrepancy between control tests proposed during class and pre-class tests proposed at the very beginning of the class (i.e., when students' attention is expected to be optimal). The authors show a strong correlation between high attention ratings obtained through their method and good performance in post-test reading comprehension.

In Wang et al., the authors examine methods for assessing people's Knowledge, Attitude, and Behavior (KAB) regarding medical AI. In doing so, they compare a person-based approach that stratifies a population's KAB based on individual profiles with the more common variable-based approach relying on isolated self-assessments of each component. This approach highlights the emergence of subtler profiles of interaction among the three components.

Overall, these papers provide valuable insights into understanding user experience, attention, and attitudes in AI interactive systems, offering new scales, assessment methods, and approaches for further exploration.

Opinion and perspective contributions

Since AI systems are increasingly relied upon for decision-making across different domains, limitations and risks associated with certain applications of AI need to be taken into consideration. Nathan and Fourneret and Yvert aim to shed light on critical issues associated with the use of AI systems.

Nathan (2023) focus on the limitations of disembodied AI (dAI) in educational systems, which emerged particularly during the COVID-19 pandemic. Such systems have two significant limitations: they struggle to model people's embodied interactions, as they primarily rely on statistical regularities rather than capturing the nuanced nature of human behavior; and they are often black boxes, lacking transparency and predictability when applied to new domains. The emergence of multimodal learning analytics and data mining (MMLA) exacerbates the issue, as data accessibility and usage are not properly regulated. To mitigate the risks associated with dAI, Nathan proposes an alternative augmented intelligence system that effectively addresses students' needs.

On the other hand, Fourneret and Yvert highlight a more subtle risk associated with using AI systems to aid human decision-making: human desubjectivation. People's increasing reliance on AI system recommendations has led to various forms of digital normativity, where algorithms establish standards that individuals adopt as the norm in their daily lives, a phenomenon that may affect the acquisition and exercise of subjectivity, influencing critical thinking. Relying entirely on AI systems for decision-making promotes human comfort but discourages individuals from challenging or refusing system suggestions due to their perceived infallibility. To address the risk of desubjectivation, Fourneret and Yvert highlight the importance of an Ethics-by-design methodology, involving ways to protect the subjective thinking process during the project's ideation phase rather than at implementation. They emphasize the importance of involving philosophers and ethicists in the development of new technologies and emphasize the need to educate future generations about the risks of silent acceptance of AI governmentality (see also Franzoni, 2023).

Open problems and future work

Despite the ongoing debates and discussions regarding the ethical aspects of AI, practical solutions to ensure shared ethics remain open challenges.

Transparency and explainability

One of the significant challenges lies in the transparency and explainability of AI systems. Generalist AI systems often employ sophisticated algorithms and deep neural networks, making it difficult to understand and explain their decision-making processes (see Adadi and Berrada, 2018; Balasubramaniam et al., 2023). The lack of transparency and interpretability raises concerns about discrimination and unfair or unjust outcomes.

Accountability and responsibility, autonomy, human oversight, and control

As AI systems take on increasingly autonomous decision-making roles, traditional models of responsibility may not adequately capture the new unique challenges posed. Establishing clear frameworks for assigning responsibility and addressing questions of negligence, oversight, and the potential for unintended consequences is essential to ensure accountability for the decisions made by AI systems, capable of making autonomous decisions across various domains without human intervention. Balancing the autonomy of AI systems with human judgment and intervention is necessary to prevent undue reliance on AI decisions and preserve human agency and accountability (see Beckers, 2022; Cavalcante Siebert and Lupetti, 2023).

Bias and fairness, Societal impact, and distribution of benefits

AI systems can inadvertently perpetuate biases present in the data they are trained on, leading to discriminatory outcomes (see Dwork, 2012; Mehrabi, 2021). Decision-making AI must be designed to recognize and mitigate biases, ensuring fairness in the decision-making process across diverse populations. This issue requires developing techniques that identify and address biases and allow designers to be conscious of their biases and limits. AI systems can have significant societal impacts, influencing resource allocation, access to services, and opportunities. Ensuring these systems are designed and deployed to benefit all individuals is critical to avoid exacerbating existing inequalities (Datta, 2023).

Privacy and data security

Generalist AI systems rely on vast amounts of data, often including sensitive personal information. Protecting individual privacy and ensuring robust data security measures become paramount to prevent misuse or unauthorized access to personal information. Balancing the benefits of artificial intelligence with privacy considerations is an ongoing challenge, as a huge number of entities are massively and continuously collecting data, virtually beyond the control of individuals (see Song et al., 2022; USA White House Executive Office, 2023).

Conclusion

In the new era of Generalist AI, where AI systems are expected to handle a wide range of tasks and exhibit human-like capabilities, the challenges, and complexities of ethical design will become more pronounced. AI systems may encounter situations where ethical dilemmas arise, such as conflicts between different moral values or competing interests (see Xiaoling, 2021; Huang et al., 2022). Deciding how to prioritize and navigate these ethical dilemmas becomes crucial. Establishing clear ethical frameworks and guidelines to make ethically sound decisions is a complex challenge (see Ramos and Koukku-Ronde, 2022; UNESCO, 2022). Researchers must explore interdisciplinary collaborations that combine expertise in AI, ethics, philosophy, law, and social sciences. This collaborative approach can pave the way for developing comprehensive ethical frameworks, and standards that govern the design, deployment, and use of AI-based decision-making systems.

Author contributions

VF and JV: conception and design of the work. JV: draft—Sections Introduction and State of the art. GB: draft—Section Systematic reviews. SC: draft—Section Novel research contributions. RC: draft—Section Opinion and perspective contributions. AM: draft—Sections Open problems and future work, Conclusion. VF and FL: critical revision of the manuscript. VF: work supervision and review. All authors provide approval for publication of the content and agree to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

VF, AM, and GB are supported by the EmoRe Research Group of the University of Perugia. JV was supported by an ICREA Acadèmia Research Grant ICREA2019. In this work, FL was partially supported by the project FAIR—Future AI Research (PE00000013), spoke 6—Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., and Kujala, S. (2023). Transparency and explainability of AI systems: from ethical guidelines to requirements. *Inform. Softw. Technol.* 159, 107197. doi: 10.1016/j.infsof.2023.107197

Beckers, N. (2022). Drivers of partially automated vehicles are blamed for crashes that they cannot reasonably avoid. *Sci. Rep.* 12, 16193. doi: 10.1038/s41598-022-19876-0

Biondi, G., Franzoni, V., Mancinelli, A., Milani, A., and Niyogi, R. (2022). "Hate speech and stereotypes with artificial neural networks," in *Computational Science and Its Applications – ICCSA 202* (Malaga).

Biondi, G., Franzoni, V., and Milani, A. (2022). "Defining classification ambiguity to discover a potential bias applied to emotion recognition data sets," in *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 672–679.

Bryson, J., and Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer* 50, 116–119. doi: 10.1109/MC.2017.154

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency* (PMLR), 77–91.

Cavalcante Siebert, L., and Lupetti, M. L. (2023). Meaningful human control: actionable properties for AI system development. *AI Ethics* 3, 241–255.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci. Eng. Ethics* 26, 2051–2068. doi: 10.1007/s11948-019-00146-8

Crawford, K., and Calo, R. (2016). There is a blind spot in AI research. *Nature* 538, 311–313. doi: 10.1038/538311a

Datta, T. (2023). "Tensions between the proxies of human values in AI," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, (New York, NY, United States) 678–689.

Dwork, C. (2012). "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12* (New York, NY: Association for Computing Machinery), 214–226.

European Commission (2019). "High-level expert group on artificial intelligence," in *Ethics Guidelines for Trustworthy AI* (European Commission), 6.

Floridi, L., Cowls, J., King, T. C., and Taddeo, M. (2020). How to design AI for social good: seven essential factors. *Sci. Eng. Ethics* 26, 1771–1796. doi: 10.1007/s11948-020-00213-5

Franzoni, V. (2023). "From black box to glass box: advancing transparency in artificial intelligence systems for ethical and trustworthy AI," in *Computational Science and Its Applications–ICCSA 2023* (Athens: Springer).

Franzoni, V., and Milani, A. (2019). "Emotion recognition for self-aid in addiction treatment, psychotherapy, and nonviolent communication," in *Computational Science and Its Applications–ICCSA 2019: 19th International Conference* (St. Petersburg: Springer), 391–404.

Huang, C., Zhang, Z., Mao, B., and Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Trans. Artif. Intell.* 1–21. doi: 10.1109/TAI.2022.3194503

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1, 389–399. doi: 10.1038/s42256-019-0088-2

Larsson, S., and Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Rev.* 9. doi: 10.14763/2020.2.1469

Madiega, T. A. (2021). *Artificial Intelligence Act*. European Parliament: European Parliamentary Research Service.

Manheim, K., and Kaplan, L. (2019). Artificial intelligence: risks to privacy and democracy. *Yale JL Tech.* 21, 106.

Mehrabi, N. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54. doi: 10.1145/3457607

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* 1, 501–507. doi: 10.1038/s42256-019-0114-4

Nathan, M. J. (2023). Disembodied AI and the limits to machine understanding of students' embodied interactions. *Front. Artif. Intell.* 6, 1148227. doi: 10.3389/frai.2023.1148227

Novelli, C., Taddeo, M., and Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. *AI Soc.* 1–12. doi: 10.1007/s00146-023-01635-y

Ramos, G., and Koukku-Ronde, R. (2022). UNESCO's global agreement on the ethics of AI can guide governments and companies alike.

Song, J., Han, Z., Wang, W., Chen, J., and Liu, Y. (2022). A new secure arrangement for privacy-preserving data collection. *Comput. Standards Interfaces* 80, 103582. doi: 10.1016/j.csi.2021.103582

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST.

UNESCO (2022). *Recommendation on the Ethics of Artificial Intelligence*.

USA White House Executive Office (2023). *Report: National Strategy to Advance Privacy-Preserving Data Sharing and Analytics*.

Vallverdú, J., and Casacuberta, D. (2014). "Ethical and technical aspects of emotions to create empathy in medical machines," in *Machine Medical Ethics*, eds S. P. van Rysewyk and M. Pontier (Springer), 341–362.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., et al. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nat. Commun.* 11, 233. doi: 10.1038/s41467-019-14108-y

Xiaoling, P. (2021). "Discussion on ethical dilemma caused by artificial intelligence and countermeasures," in *2021 IEEE Conference on Image Processing, Electronics and Computers* (Dalian: IPEC), 453–457.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*, eds A. Singh, and J. Zhu (Fort Lauderdale, FL: PMLR), 962–970.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 21 November 2019
doi: 10.3389/frai.2019.00024

Can I Influence You? Development of a Scale to Measure Perceived Persuasiveness and Two Studies Showing the Use of the Scale

Rosemary J. Thomas[1,2]*, *Judith Masthoff*[1,3]* and *Nir Oren*[1]*

[1] Computing Science, University of Aberdeen, Aberdeen, United Kingdom, [2] Institute of Global Food Security, Queen's University Belfast, Belfast, United Kingdom, [3] Computing Science, Utrecht University, Utrecht, Netherlands

In this paper, we develop and validate a scale to measure the perceived persuasiveness of messages to be used in digital behavior interventions. A literature review is conducted to inspire the initial scale items. The scale is developed using Exploratory and Confirmatory Factor Analysis on the data from a study with 249 ratings of healthy eating messages. The construct validity of the scale is established using ratings of 573 email security messages. Using the data from the two studies, we also show the usefulness of the scale by analyzing the perceived persuasiveness of different message types on the developed scale factors in both the healthy eating and email security domains. The results of our studies also show that the persuasiveness of message types is domain dependent and that when studying the persuasiveness of message types, the finer-grained argumentation schemes need to be considered and not just Cialdini's principles.

Keywords: perceived persuasiveness, scale development, behavior change, message type, argumentation schemes

OPEN ACCESS

Edited by:
Ralf Klamma,
RWTH Aachen University, Germany

Reviewed by:
Jorge Luis Bacca Acosta,
Fundación Universitaria Konrad Lorenz, Colombia
Kirsten Ailsa Smith,
University of Southampton, United Kingdom

***Correspondence:**
Rosemary J. Thomas
rosemaryjthomas@acm.org
Judith Masthoff
j.f.m.masthoff@uu.nl
Nir Oren
n.oren@abdn.ac.uk

Specialty section:
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

Received: *21 May 2019*
Accepted: *25 October 2019*
Published: *21 November 2019*

Citation:
Thomas RJ, Masthoff J and Oren N (2019) Can I Influence You? Development of a Scale to Measure Perceived Persuasiveness and Two Studies Showing the Use of the Scale. Front. Artif. Intell. 2:24. doi: 10.3389/frai.2019.00024

1. INTRODUCTION

Many behavior change interventions have been developed for a wide variety of domains. For example, "Fit4Life" (Purpura et al., 2011) promotes healthy weight management, the ASICA application (Smith et al., 2016) reminds skin-cancer patients to self-examine their skin, the SUPERHUB application (Wells et al., 2014) motivates sustainable travel, while "Portia" (Mazzotta et al., 2007) and "Daphne" (Grasso et al., 2000) encourage healthy eating habits.

Clearly, it is important to measure the effectiveness of such persuasive interventions. However, it is often difficult to measure actual persuasiveness (O'Keefe, 2018). Perhaps the primary three reasons for such difficulties are as follows. First, measuring actual persuasiveness tends to require more time and effort from participants and additional resources. For example, to measure the persuasiveness of a healthy eating intervention, participants may need to provide detailed diaries of their food intake, which are cumbersome and often unreliable (Cook et al., 2000), and may require the provisioning of scales to participants. Also, when studying many experimental conditions, it may be hard to obtain sufficient participants willing to spend the necessary time [e.g., to measure actual persuasiveness of reminders in (Smith et al., 2016) would have required a large number of skin cancer patients]. Second, it is hard to measure actual persuasiveness due to confounding factors. For example, when measuring the persuasiveness of a sustainable transport application,

other factors such as the weather may influence people's behavior. Third, there may be ethical issues which make it hard to measure actual persuasiveness. For example, if one wanted to investigate the persuasive effects of different message types to get learners to study more, it may be deemed unethical to do this in a real class room, as learners in the control condition may be seen to be disadvantaged. Purpura et al. (2011) illustrates some of the ethical problems while using persuasive technologies in behavior change interventions.

Because of these difficulties in measuring actual persuasiveness, *perceived* persuasiveness is often used as an approximation of, or the initial step in the measurement of, actual persuasiveness (see **Table 1** for example studies that used perceived persuasiveness). Perceived persuasiveness may include multiple factors. For example, perceived effectiveness in changing somebody's attitudes may be different from perceived effectiveness in changing behavior. We would like a reliable scale that incorporates multiple factors as sub-scales, with each sub-scale consisting of multiple items. Such a scale does not yet exist, and researchers have so far had to use their own measures without proper validation.

Therefore, this paper describes the process for developing a reliable and validated multi-item, multi-subscale scale to measure perceived persuasiveness. In addition, the data collected will be used to show the usefulness of the scale by analyzing the impact of different persuasive message types on the developed scale factors.

2. LITERATURE REVIEW

To inspire the scale items and show the need for scale development, we first investigated how researchers measured perceived persuasiveness by examining the scale items and respective measurements they used in published user studies. We performed a semi-structured literature review, searching in Scopus from the period 2014 to 2018 across disciplines. At first we performed a narrow search using the following search query:

"scales development" AND studies AND persuasion.

However, this produced very few search results. Later, we modified the search query to the following:

persuasion AND (experiments OR studies)

to get a broader range of articles. We also searched in the Proceedings of the "International Conference on Persuasive Technology" for the period from 2013 to 2018. We were looking for user studies that developed or used a scale to measure perceived persuasiveness. The search resulted in 12 papers, including 2 from outside computer science from marketing and communications (Koch and Zerbac, 2013; Zhang et al., 2014). Ham et al. (2015) and O'Keefe (2018) appeared in the initial search results but were excluded as they contained meta-reviews rather than original studies. Three papers were added to the results through snowballing, given these specifically addressed perceived persuasiveness scales:

- Kaptein et al. (2009) cited in Busch et al. (2013).
- MacKenzie and Lutz (1989) cited in Ham et al. (2015).
- Zhao et al. (2011) cited in O'Keefe (2018).

The results of the literature search are shown in **Table 1**, which lists 60 scale items and their measurements based on studies reported in these 15 papers[1].

Unfortunately, most studies do not report on the scale construction, reliability or validation. The exceptions are Kaptein et al. (2009) and Busch et al. (2013). However, Kaptein et al. (2009)'s scale really measures the susceptibility of participants to certain Cialdini's principles of persuasion (such as liking and authority) (Cialdini, 2009), rather than the persuasion of the messages themselves. Similarly, Busch et al. (2013) aims to measure the persuasibility of participants by certain persuasive strategies (such as social comparison and rewards).

We reduced the 60 items listed in **Table 1** in two steps. First, we removed duplicates and merged highly similar items. Next, we transformed items that were not yet related to a message where possible (items 9, 11–13, 35–36). For instance, item 11 "This feature would make me more aware of [policy]" was changed into "This message makes me more aware of my behavior," and item 35 "I always follow advice from my general practitioner" was changed into "I will follow this message." Finally, we removed items for which this was not possible (e.g., items 37–44 that measure a person's susceptibility, and items such as 10, 55). This reduced the list to the 30 items used for the initial scale development as shown in **Table 2**, which also shows which original items these were derived from.

A limitation of our systematic literature review is that it was mainly restricted to papers published in the period 2014–2018[2]. Additionally, it is possible for a systematic review to miss papers due to the search terms used or the limitation of searching abstracts, titles, and keywords. Some other papers related to measuring persuasiveness were found after the review was completed, most noticeably (Feltham, 1994; Allen et al., 2000; Lehto et al., 2012; Popova et al., 2014; Jasek et al., 2015; Yzer et al., 2015; McLean et al., 2016). We will discuss how the scales developed in this paper relate to this other work in our discussion section.

3. STUDY DESIGN

3.1. Study 1: Development of a Perceived Persuasiveness Scale

We conducted a study to develop a rating scale to measure the "perceived persuasiveness" of messages. The aim was to obtain a scale with good internal consistency, and with at least three items per factor following the advice in MacCallum et al. (1999) to have at least three or four items with high loadings per factor.

[1]Many of these papers contained additional items; these were normally not related to measuring persuasiveness.

[2]In this period much research on persuasive technology has taken place, as evidenced by 7,410 papers being found for "persuasive technology" in Google Scholar.

TABLE 1 | Scale items related to measuring perceived persuasiveness, the measurement scale used for each item, and the number of measurement points.

References	Number	Scale items	Scale measurement	Points
Anagnostopoulou et al. (2017)		The [System] would	Strongly disagree to strongly agree	7
	1	Influence me		
	2	be convincing		
	3	be personally relevant for me		
	4	make me [target behavior]		
Thomas et al. (2017)	5	Motivational	Not very motivating to very motivating	5
	6	Appropriateness	Very inappropriate to very appropriate	
	7	Effectiveness	Very ineffective to very effective	
	8	Convincing	Very unconvincing to very convincing	
Busch et al. (2016)	9	I find this feature useful	Strongly disagree to strongly agree	7
	10	I enjoy using this feature		
		This feature would		
	11	make me more aware of [policy]		
	12	have a positive influence on my attitude toward [policy]		
	13	lead me to comply with [policy].		
Oduor and Oinas-Kukkonen (2017)		The system provides	Strongly disagree to strongly agree	7
	14	trustworthy content		
	15	believable content		
	16	accurate content		
	17	professional information		
Chang et al. (2018)		On average, [communications] are	Strongly disagree to strongly agree	5
	18	persuasive		
	19	compelling		
	20	logical		
	21	plausible		
Zhao et al. (2011)		[Communication] that is	Strongly disagree to strongly agree	5
	22	believable		
	23	convincing		
	24	important to me		
Orji (2014), Orji et al. (2014)		The system would	Strongly disagree to strongly agree	7
	25	influence me		
	26	be convincing		
	27	be personally relevant for me		
	28	make me reconsider my [behavior]		
Zhang et al. (2014)		[Communications] were	Strongly disagree to strongly agree	7
	29	convincing		
	30	persuasive		
	31	strong		
	32	good		
	33	trustworthy		
	34	reliable		
Kaptein et al. (2009)[a]		Susceptibility authority:	Totally disagree to Totally agree	7
	35	I always follow advice from my general practitioner.		
	36	When a professor tells me something I tend to believe it is true.		
		Susceptibility consensus:		
	37	If someone from my social network notifies me about a good book, I tend to read it.		

(Continued)

TABLE 1 | Continued

References	Number	Scale items	Scale measurement	Points
	38	When I am in a new situation I look at others to see what I should do.		
		Susceptibility liking:		
	39	I accept advice from my social network.		
	40	When I like someone, I am more inclined to believe him or her.		
Busch et al. (2013)	41	Before I do something, I want to know how other people have done it, so I can feel more safe.	Fully agree to Fully disagree	9
	42	It is important to me to know what other people are doing.		
	43	I trust information better where the source is specified.		
	44	It is important for me to be precisely informed about things that I need to do, before I do them.		
Hammer et al. (2016)	45	[Communications were]	Not polite to Very polite Not persuasive at all to Very persuasive	7
Hossain and Saini (2014)	46	The [communication] is	Truthful to Not truthful Unbelievable to Believable Not deceptive to Deceptive	8
	47	The [communicator] is	Sincere to Insincere Honest to Dishonest Not manipulative to Manipulative Not pushy to Pushy	
Meschtscherjakov et al. (2016)		This system	Completely disagree to completely agree	7
	48	makes people change their behavior		
	49	has the potential to influence people		
	50	gives the behavior of its users a new direction		
	51	is exactly what I need to change my attitude		
	52	does not cause a change in behavior with me		
	53	causes me to do some things differently		
	54	Thanks to the system I reach my goals.		
	55	I will use this system as often as possible.		
	56	With the help of the system, I will behave differently in the future.		
Koch and Zerbac (2013)	57	I had the feeling that [communicator] wanted to convince the reader of [communicator]'s standpoint	I do not agree at all to I fully agree	5
	58	[Communicator] wanted to convince me of [communicator]'s views		
MacKenzie and Lutz (1989)	59	Attitude: The [communicator/communication] is	Good to bad Pleasant to Unpleasant Favorable to Unfavorable	7
	60	Credibility: The [communicator/communication] is	Convincing to Unconvincing Believable to Unbelievable Biased to Unbiased	

[a] *This is a sample. They also present items related to the susceptibility to the other Cialdini principles.*

3.1.1. Participants

The participants for this study were recruited by sharing the link of the study via social media and mailing lists. The study had four validation questions to check if participants were randomly rating the scales. After removing such participants, a total of 92 participants rated 249 messages.

3.1.2. Procedure

Each participant was shown a set of five messages (see **Table 4**), each promoting healthy eating. These messages were based on different *argumentation schemes*[3] (Walton et al., 2008) and were produced in another study using a message generation system (Thomas et al., 2018). Each message was rated using 34 scale items (the scale items marked with * act as validation checks) on a 7-point Likert scale that ranges from "strongly disagree" to "strongly agree" (see **Table 2** and **Figure 1**). Finally, participants were given the option to provide feedback.

[3] Argumentation schemes are stereotypical patterns of reasoning.

TABLE 2 | Scale items developed used in Study 1.

Scale items		Inspired from / Similar to
This message is	influencing.®	1,25
	convincing. ®	2,8,23,26,29,57,58
	personally relevant.®	3,24,27
	motivating. ®	5
	appropriate.	6
	credible.	21,31,60
	encouraging. ®	7,18,30,45
	inappropriate.*	N/A
	effective.®	7
	useful.®	9,17,32
	believable. ®	15,22,34,46,47
	ineffective.*	N/A
	accurate.	16,17,20
	trustworthy.	14,33,46,47
	exactly what I need to help reach my goals.	54
	exactly what I need to change my attitude.	49,51
	exactly what I need to change my behavior.	51
This message	makes me more aware of my behavior.®	11
	leads me to comply with behavior expectations.®	13,19
	will cause changes in my behavior.	4
	has a positive influence on my attitude.®	12
	has the potential to change user behavior.	48
	has the potential to influence user behavior.	49
	has the potential to inspire users.	50
	causes a change in my behavior.	4,52
	causes me to make some changes in my behavior.	4,53
I	will follow this message.®	13,19,35
	consider this message.	28
	accept this message.	39
	believe this message is true.®	36
After viewing this message, I will make	some behavior change in the future.	56
	changes in my attitude.	49,51
Please click the second option from the	right.*	N/A
	left.*	N/A

*act as validation check. ®cross loaded on different factors.

TABLE 3 | Study 1: Reduced scales items after EFA.

Factors	Scale items
Effectiveness	This message will cause changes in my behavior.®
	This message is exactly what I need to help reach my goals.®
	This message is exactly what I need to change my attitude.®
	This message is exactly what I need to change my behavior.®
	This message causes a change in my behavior.
	This message causes me to make some changes in my behavior.
	After viewing this message, I will make some behavior change in the future.®
	After viewing this message, I will make changes in my attitude.
Quality	This message is appropriate.®
	This message is credible.®
	This message is believable.
	This message is accurate.
	This message is trustworthy.
	I believe this message is true.
	I accept this message.®
Capability	This message has the potential to change user behavior.
	This message has the potential to influence user behavior.
	This message has the potential to inspire users.

® *high Standardized Residual Covariances with several other items.*

TABLE 4 | Healthy eating messages used in Study 1 with corresponding argumentation schemes.

Scheme name	Message
Argument from commitment with goal	As you want to eat healthy, you are committed to eating healthy foods. So, you are also committed to shopping carefully and reading the labels as it helps you to eat healthy foods.
Argument from expert opinion with goal	A nutritionist recommends that you keep a log of your daily calorie intake to manage calorie intake. So, you should follow their recommendation.
Argument from position to know with goal	A college football star suggests that you eat a diet high in protein to have more energy. So, you should follow their suggestion.
Argument from sunk cost with action	You have a choice whether or not to eat vegetables with every serving, however, you committed to doing so earlier. So, you should choose to eat vegetables with every serving.
Practical reasoning with goal	If you cut out added sugars, white flours, white rice and soft drinks, it helps you to lose weight. So, you ought to do this.

3.1.3. Research Question and Hypothesis

We were interested in the following research question:

- RQ1: What is a reliable scale to measure perceived persuasiveness?

In addition, we wanted to investigate the usefulness of the scale by analyzing whether the different message types had an impact on the ratings of the developed factors. Therefore, we formulated the following hypothesis:

- H1: Perceived persuasiveness of each factor differs for different message types.

3.2. Study 2: Validation of the Perceived Persuasiveness Scale

Next, we conducted a study to determine the construct validity of the developed scale. We replicated the scale-testing in the domain of email security using another data set.

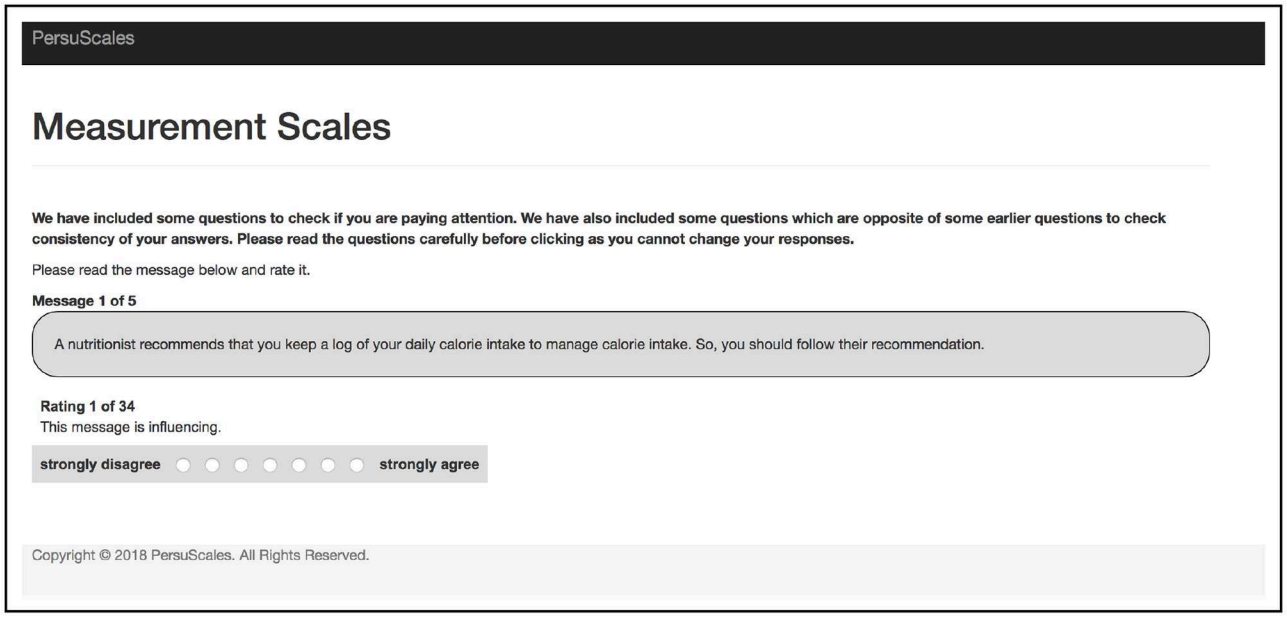

FIGURE 1 | Screenshot of Study 1 showing a message with a scale item to be rated.

TABLE 5 | Email security messages used in Study 2 with corresponding argumentation schemes.

Scheme name	Message
Argument from commitment with goal	As you want to keep your computer account safe, you are committed to check whether website links are genuine in emails. So, you are also committed to preview website links in an email application as it helps you to check whether website links are genuine in emails.
Argument from expert opinion with goal	A renowned email security expert recommends that you prevent opening suspicious attachments to protect your email account. So, you should follow their recommendation.
Argument from position to know with goal	A colleague who attended email security training suggests that you verify the logo, header and footer of email newsletters to make sure that they originate from genuine sources. So, you should follow their suggestion.
Argument from sunk cost with action	You have a choice whether or not to be security-conscious when processing email; however, you committed to doing so earlier. So, you should choose to be security-conscious when processing email.
Practical reasoning with goal	If you choose not to provide personal information by responding to emails that threaten to disable account access, it helps you to safeguard your email account. So, you ought to do this.

3.2.1. Participants

The participants for this study were recruited by sharing the link of the study via social media and mailing lists. After removing the invalid participants (as before), a total of 134 participants rated 573 messages.

TABLE 6 | Study 1: Reduced scale items after CFA.

Factors	Scale items
Effectiveness	This message will cause changes in my behavior. This message causes me to make some changes in my behavior. After viewing this message, I will make changes in my attitude.
Quality	This message is accurate. This message is trustworthy. I believe this message is true.
Capability	This message has the potential to change user behavior. This message has the potential to influence user behavior. This message has the potential to inspire users.

3.2.2. Procedure

Each participant was shown a set of five messages (see **Table 5**) that promote email security, again based on argumentation-schemes. Each message was rated using the scale (see **Table 6** and **Figure 2**) that resulted from Study 1. Finally, participants were given the option to provide feedback.

3.2.3. Research Question and Hypotheses

We were interested in the following research question:

- RQ2: How valid is the developed perceived persuasiveness scale?

Our first study: Development of a Perceived Persuasiveness Scale resulted in a scale with three factors for measuring

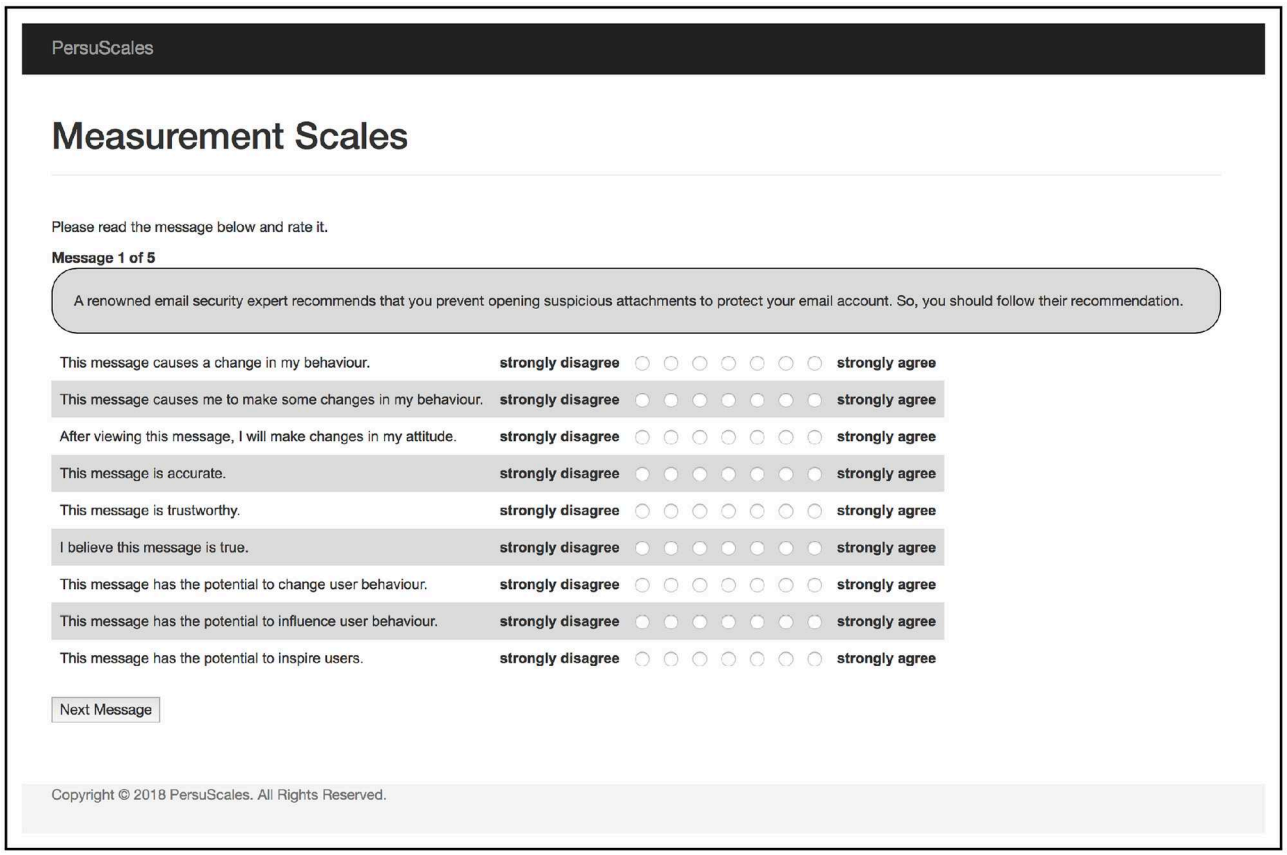

FIGURE 2 | Screenshot of Study 2 showing message with the scale items to be rated.

perceived persuasiveness: Effectiveness, Quality, and Capability (see section 4.1). We wanted to investigate the usefulness of this scale by analyzing whether the message types differed on these three developed factors. Therefore, we formulated the following hypotheses:

- H2: The perceived persuasiveness factor Effectiveness differs for different message types.
- H3: The perceived persuasiveness factor Quality differs for different message types.
- H4: The perceived persuasiveness factor Capability differs for different message types.
- H5: Overall perceived persuasiveness[4] differs for different message types.

4. RESULTS

4.1. Study 1: Development of a Perceived Persuasiveness Scale

First we checked the Kaiser-Meyer-Olkin Measure of Sampling Adequacy, which was greater than 0.90. According to this measure, values in the 0.90's indicate that the sampling adequacy is "marvelous" (Dziuban and Shirkey, 1980). Next, we investigated the inter-item correlations. For the factor analysis, all the 7-point scale items were considered as ordinal measures. To further filter the items and identify the factors, we conducted an Exploratory Factor Analysis (EFA) using Principal Component Analysis extraction and Varimax rotation with Kaiser Normalization (Howitt and Cramer, 2014). Varimax rotation was used as the matrix was confirmed orthogonal (the Component Correlation Matrix shows that the majority of the correlations was less than 0.5). We obtained three factors (see **Table 2**). The first factor we named Effectiveness as its items relate to user behavior and attitude changes and attainment of user goals. The second we named Quality as its items relate to characteristics of a message strength such as trustworthiness and appropriateness. The third we named Capability as its items relate to the *potential* for motivating users to change behavior. We removed the 13 items that cross loaded on different factors (see **Table 2** with scale items marked ®). This resulted in **Table 3**, which shows the reduced scale items for the three factors. We checked the Cronbach's Alpha of all the items belonging to the three factors separately. It was greater than 0.9 for each of the three factors which indicates "excellent" scale reliability.

[4]Overall perceived persuasiveness was calculated as the mean of the factors: Effectiveness, Quality, and Capability.

Next, we conducted Confirmatory Factor Analyses (CFA) to determine the validity of the scale, and to confirm the factors and items by checking the model fit (Hu and Bentler, 1999). Based on these analyses, 8 items were removed due to high Standardized Residual Covariances with several other items which were greater than 0.4. The items removed are the items in **Table 3** marked ®.

Table 6 shows the resulting scale of 9 items. The final Confirmatory Factor Analysis resulted in the following values for the Tucker-Lewis Index (TLI) = 0.988, Comparative Fit Index (CFI) = 0.993, and Root Mean Square Error of Approximation (RMSEA) = 0.054, when extracting the three factors and their items. A cut off value nearing 0.95 for TLI and CFI (the higher the better) and a cut off value nearing 0.60 for RMSEA (the lower the better) are required to establish that there is an acceptable model fit between the hypothesized model and the observed data (Hu and Bentler, 1999; Schreiber et al., 2006). In the resulting scale, the TLI and CFI are above 0.95 and RMSEA is below 0.60, which shows an acceptable model fit. This answers research question RQ1.

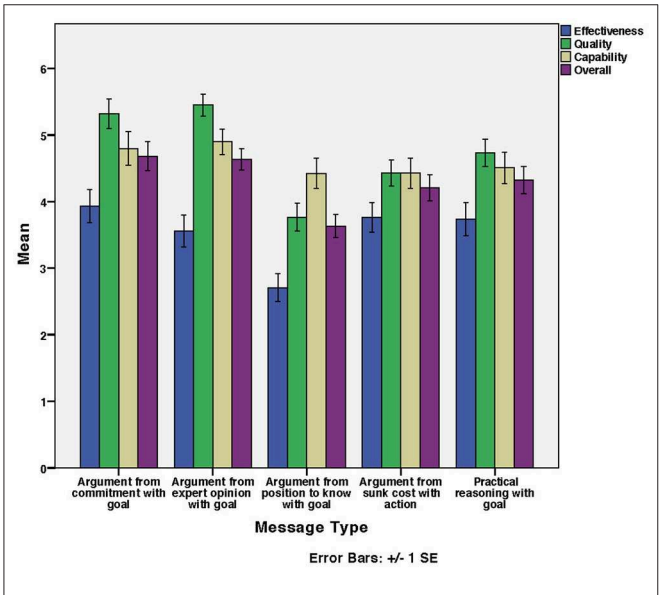

FIGURE 3 | Healthy eating messages: Mean of factors' and overall ratings for developed scale per message type.

4.2. Study 1: Impact of Message Types on Factors

Figure 3 shows the mean Effectiveness, Quality, Capability, and Overall perceived persuasiveness of message types used for the healthy eating messages. Overall perceived persuasiveness was calculated as the mean of the factors: Effectiveness, Quality, and Capability.

A one-way repeated measures MANOVA with Effectiveness, Quality, Capability, and Overall perceived persuasiveness as dependent variables and message type as the independent variable provided the results for the analyses given below. To determine the homogeneous subsets, the Ryan-Einot-Gabriel-Welsch Range was selected as a *post-hoc* test since we have more than 3 levels within the independent variable (i.e., the message type).

According to Thomas et al. (2018), the argumentation schemes can be mapped to Cialdini's principles of persuasion.

1. Cialdini's Principle: Commitments and Consistency
   Argument from commitment with goal
   Practical reasoning with goal.
   Argument from sunk cost with action
2. Cialdini's Principle: Authority
   Argument from expert opinion with goal
   Argument from position to know with goal.

The study conducted by Thomas et al. (2017) states that Authority was significantly more persuasive, followed by Commitments and Consistency and the other Cialdini principles. We were interested to know whether our findings would be similar. Hence, the analysis will consider both the argumentation schemes and Cialdini's principles when discussing the findings.

4.2.1. Impact of Message Types on Effectiveness

According to **Figure 3**, ARGUMENT FROM COMMITMENT WITH GOAL was the highest rated in Effectiveness while ARGUMENT FROM POSITION TO KNOW WITH GOAL was the lowest. There was a significant effect of message type on Effectiveness [$F_{(4, 244)} = 4.39$, $p < 0.01$]. There was a significant difference between ARGUMENT FROM POSITION TO KNOW WITH GOAL and the other message types ($p < 0.05$). The rest were non-significant. **Table 7** shows the homogeneous subsets. This partially supports the hypothesis (H1) that perceived persuasiveness on each factor differs for different message types.

TABLE 7 | Study 1: Homogeneous subsets for Effectiveness, Quality, and Capability.

Message type	Effectiveness			Quality					Capability	
	Mean			Mean						Mean
	N	S1	S2	N	S1	S2	S3	S4	N	S1
Argument from position to know with goal	52	2.71		52	3.76				52	4.42
Argument from expert opinion with goal	52		3.56	51		4.43			51	4.42
Practical reasoning with goal	48		3.74	48		4.73	4.73		48	4.51
Argument from sunk cost with action	51		3.76	46			5.32	5.32	46	4.80
Argument from commitment with goal	46		3.93	52				5.45	52	4.90

TABLE 8 | Study 1: Homogeneous subsets for Overall Perceived Persuasiveness.

Message type	*N*	Mean	
		Subset 1	Subset 2
Argument from position to know with goal	52	3.63	
Argument from sunk cost with action	51	4.21	4.21
Practical reasoning with goal	48	4.32	4.32
Argument from expert opinion with goal	52		4.63
Argument from commitment with goal	46		4.68

As shown, the two Authority messages had the lowest Effectiveness scores, though the ARGUMENT FROM EXPERT OPINION WITH GOAL was not rated significantly lower than the Commitments and Consistency messages. We observe that the Effectiveness of all messages was low, below or around the mid-point of the scale. This contradicts the results from Thomas et al. (2017) where Authority and Commitments and Consistency messages were most persuasive, though of course their study only considered overall perceived persuasiveness without using a validated scale.

4.2.2. Impact of Message Types on Quality

According to **Figure 3**, for healthy eating messages ARGUMENT FROM EXPERT OPINION WITH GOAL was the highest rated in quality while ARGUMENT FROM POSITION TO KNOW WITH GOAL was the lowest. There was a significant effect of message type on Quality [$F_{(4, 244)} = 12.14$, $p < 0.001$]. There was a significant difference ($p < 0.05$) between:

1. ARGUMENT FROM POSITION TO KNOW WITH GOAL and the other message types,
2. ARGUMENT FROM SUNK COST WITH ACTION and the other message types except PRACTICAL REASONING WITH GOAL,
3. PRACTICAL REASONING WITH GOAL and the other message types except ARGUMENT FROM COMMITMENT WITH GOAL, and
4. ARGUMENT FROM COMMITMENT WITH GOAL and the other message types except ARGUMENT FROM EXPERT OPINION WITH GOAL.

Table 7 shows the homogeneous subsets. This partially supports the hypothesis (H1) that perceived persuasiveness on each factor differs for different message types. However, it should be noted that one Authority message is the worst and one the best on Quality. This may either be caused by attributes of the message itself, or by one of the Authority argumentation schemes resulting in higher quality messages than the other one.

4.2.3. Impact of Message Types on Capability

According to **Figure 3**, ARGUMENT FROM EXPERT OPINION WITH GOAL was slightly higher rated in quality compared to the other message types. There was no significant effect of message type on Capability [$F_{(4, 244)} = 0.98$, $p > 0.05$]. **Table 7** shows the homogeneous subsets. This does not support the hypothesis (H1) that perceived persuasiveness of each factor differs for different message types. All message types performed equally well on Capability, which was above the midpoint of the scale.

4.2.4. Impact of Message Types on Overall Perceived Persuasiveness

According to **Figure 3**, ARGUMENT FROM COMMITMENT WITH GOAL was the highest rated overall while ARGUMENT FROM POSITION TO KNOW WITH GOAL was the lowest. There was a significant effect of message type on Overall Perceived Persuasiveness [$F_{(4, 244)} = 4.98$, $p < 0.01$]. ARGUMENT FROM POSITION TO KNOW WITH GOAL was significantly different from ARGUMENT FROM EXPERT OPINION WITH GOAL and ARGUMENT FROM COMMITMENT WITH GOAL ($p < 0.05$). The rest were non-significant. **Table 8** shows the homogeneous subsets. This partially supports the hypothesis (H1) that each factor differs on different message types.

4.3. Study 2: Validation of the Perceived Persuasiveness Scale

To determine the construct validity of the developed scale in Study 1 and replicate the scale-testing, we:

1. Used an 80-20 split validation on the original dataset of Study 1. With this specific combination, the developed scale resulted in an acceptable model fit for 80% (TLI = 0.975, CFI = 0.985, RMSEA = 0.081) and 20% of the data (TLI = 0.975, CFI = 0.985, RMSEA = 0.080).
2. Used the dataset obtained from the validation in Study 2. With this dataset, the developed model resulted in an acceptable fit (TLI = 0.984, CFI = 0.990, RMSEA = 0.071).

This answers research question RQ2, validating the scale.

4.4. Study 2: Impact of Message Types on Factors

Figure 4 shows the mean Effectiveness, Quality, Capability, and Overall perceived persuasiveness of message types used for email security messages. As before, the Overall perceived persuasiveness was calculated as the mean of the factors Effectiveness, Quality, and Capability.

A one-way repeated measures MANOVA with Effectiveness, Quality, Capability, and Overall perceived persuasiveness as dependent variables and message type as the independent variable provided the results for the analyses given below. To determine the homogeneous subsets, the Ryan-Einot-Gabriel-Welsch Range was selected as *post-hoc* test since we have more than 3 levels within the independent variable (i.e., message type).

4.4.1. Impact of Message Types on Effectiveness

According to **Figure 4**, ARGUMENT FROM EXPERT OPINION WITH GOAL was the highest rated in Effectiveness while ARGUMENT FROM COMMITMENT WITH GOAL was the lowest. There was a significant effect of message type on Effectiveness [$F_{(4, 568)} = 4.77$, $p < 0.01$]. ARGUMENT FROM COMMITMENT WITH GOAL was significantly different from ARGUMENT FROM POSITION TO KNOW WITH GOAL and ARGUMENT FROM EXPERT OPINION WITH GOAL ($p < 0.05$). The rest were non-significant. **Table 9** shows the homogeneous subsets. This partly

supports hypothesis H2, namely that perceived persuasiveness in terms of Effectiveness differs for different message types.

The subsets show that Authority messages in the email security domain performed better on Effectiveness than Commitments and Consistency messages. This is in line with the findings of the study by Thomas et al. (2017) and contradicts what was found in Study 1 for the healthy eating messages.

4.4.2. Impact of Message Types on Quality

According to **Figure 4**, ARGUMENT FROM EXPERT OPINION WITH GOAL was the highest rated in Quality while ARGUMENT FROM COMMITMENT WITH GOAL was the lowest. There was a significant effect of message type on Quality [$F_{(4, 568)} = 11.97$, $p < 0.001$]. ARGUMENT FROM EXPERT OPINION WITH GOAL was significantly different from the other message types ($p < 0.05$). The rest were non-significant. **Table 9** shows the homogeneous subsets. This partially supports hypothesis H3, namely that perceived persuasiveness in terms of Quality differs for different message types.

We observe that ARGUMENT FROM EXPERT OPINION WITH GOAL was rated significantly higher than the other message types and that the other Authority message had the second highest mean. Therefore, in the domain of email security, we can conclude that principle of Authority seems most persuasive when considering Quality. We note that ARGUMENT FROM EXPERT OPINION WITH GOAL performed best on Quality in both Studies, so this argumentation scheme seems to result in good quality messages. In contrast, ARGUMENT FROM POSITION TO KNOW WITH GOAL did not do as well in the healthy eating domain. It is possible that this is a domain effect, with people trusting people with experience more in the cyber-security domain than in the healthy eating domain. We will investigate this finding further as future work.

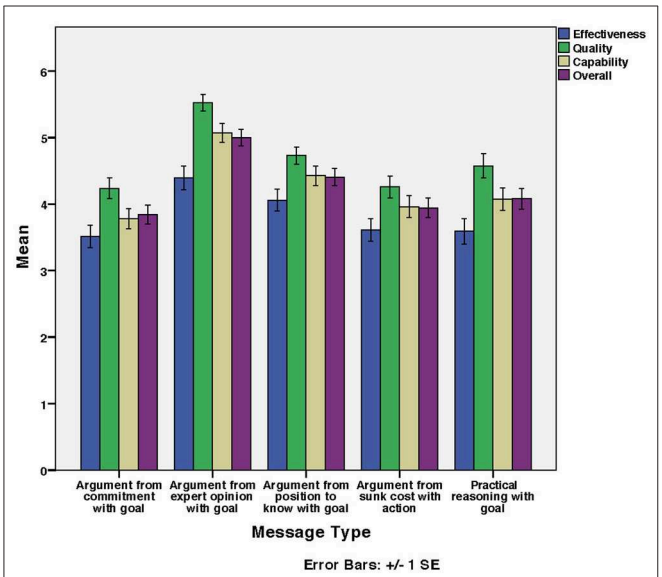

FIGURE 4 | Email security messages: Mean of factors' and overall ratings for developed scale per message type.

4.4.3. Impact of Message Types on Capability

According to **Figure 4**, ARGUMENT FROM EXPERT OPINION WITH GOAL was the highest rated in Capability while ARGUMENT FROM COMMITMENT WITH GOAL was the lowest. There was a significant effect of message type on Capability [$F_{(4, 568)} = 10.84$, $p < 0.001$]. There was significant difference ($p < 0.05$) between

1. ARGUMENT FROM EXPERT OPINION WITH GOAL and the other message types.
2. ARGUMENT FROM COMMITMENT WITH GOAL and ARGUMENT FROM POSITION TO KNOW WITH GOAL.

There were no significant differences between ARGUMENT FROM SUNK COST WITH ACTION and PRACTICAL REASONING WITH GOAL. **Table 9** shows the homogeneous subsets. This partially supports hypothesis H4 that perceived persuasiveness in terms of Capability differs for different message types.

We observe that ARGUMENT FROM EXPERT OPINION WITH GOAL was rated significantly higher than other message types, and that the other Authority message was rated second highest. Therefore, we can conclude that the principle of Authority was also most persuasive when considering Capability. Again, we can see domain effects in this finding, with ARGUMENT FROM POSITION TO KNOW performing better compared to other message types in the email security domain.

4.4.4. Impact of Message Types on Overall Perceived Persuasiveness

According to **Figure 4**, ARGUMENT FROM EXPERT OPINION WITH GOAL was the highest rated in overall perceived persuasiveness whilst ARGUMENT FROM COMMITMENT WITH

TABLE 9 | Study 2: Homogeneous subsets for Effectiveness, Quality, and Capability.

Message type	Effectiveness			Quality			Capability			
	Mean			Mean			Mean			
	N	S1	S2	N	S1	S2	N	S1	S2	S3
Argument from commitment with goal	115	3.51		115	4.24		115	3.78		
Practical reasoning with goal	111	3.59		115	4.26		115	3.96	3.96	
Argument from sunk cost with action	115	3.61		111	4.58		111	4.08	4.08	
Argument from position to know with goal	117	4.06	4.06	117	4.73		117		4.43	
Argument from expert opinion with goal	115		4.40	115		5.52	115			5.07

TABLE 10 | Study 2: Homogeneous subsets for overall perceived persuasiveness.

Message type	N	Mean		
		Subset 1	Subset 2	Subset 3
Argument from commitment with goal	115	3.84		
Argument from sunk cost with action	115	3.94	3.94	
Practical reasoning with goal	111	4.08	4.08	
Argument from position to know with goal	117		4.41	
Argument from expert opinion with goal	115			5.00

GOAL was the lowest. There was a significant effect of message type on overall perceived persuasiveness [$F_{(4, 568)} = 11.24$, $p < 0.001$]. **Table 10** shows the homogeneous subsets. This partially supports hypothesis H5 that the overall perceived persuasiveness differs for different message types.

The overall perceived persuasiveness results are similar to those for "Impact of message type on Capability"; again overall Authority messages performed well, and better than in the healthy eating domain.

5. DISCUSSION

Our studies resulted in a validated perceived persuasiveness scale as well as insights into the perceived persuasiveness of different message types.

5.1. The Perceived Persuasiveness Scale

Regarding the scale, as mentioned in the limitations of the systematic literature review, there are some other papers that proposed persuasiveness scales that were not part of the review. The uptake of these scales has been limited as judged by them not having been used in the reviewed papers. However, it is interesting to see how these scales compare to the one developed in this paper, and to consider what overlap/differences there are.

First, Feltham (1994) developed and validated a Persuasive Discourse Inventory (PDI) scale based on Aristotle's three types of persuasion: ethos, pathos, and logos (see **Table 11**). Ethos relates to the credibility of the message source, pathos to the message's affective appeal, and logos to its rational appeal. To validate the PDI scale, they mainly considered Cronbach's alpha rather than conducting a factor analysis as was done in this paper. Their results suggest that there may be cross-loadings between their scale factors as they found a positive correlation between Logos and Ethos. They also did not consider whether the scale performed well across domains, as their reassessment was conducted in a very similar domain. Regarding the scale content, the scale developed in this paper has more items that directly inquire into a message's perceived persuasiveness rather than the emotional and logical elements present in the messages, though Ethos, Logos, and Pathos still play a role. Several Ethos related items were included in our initial scale development items, namely trustworthy, believable, and credible. One of these items (cf. trustworthy) has remained in the validated scale as part of the Quality factor. The "accurate" item that is part of the Quality factor can be interpreted as on the overlap between Ethos and Logos, as it on the one hand gives a sense of being reliable, and on the other of being based on facts/rational/logical. Regarding Pathos, the item "This message has the potential to inspire users" in the Capability factor is clearly related to Pathos (as was the item "motivating" that did not make it into the final scale).

TABLE 11 | Persuasive Discourse Inventory (Feltham, 1994).

Ethos sale items: Ethos = E1+E2+E3+E4+E5 (range: 5-35)
E1) unbelievable / believable
E2) not credible / credible
E3) not trustworthy / trustworthy
E4) unreliable / reliable
E5) undependable / dependable
Logos scale items: Logos = LI+12+L3+14+L5 (range 5-35)
L1) not rational / rational
L2) not informative / informative
L3) does not deal with facts / deals with facts
L4) not knowledgeable / knowledgeable
L5) not logical / logical
Pathos scale items: Pathos = PI+P2+P3+P4+P5+P6+P7 (range: 7-49)
P1) does not affect my feelings / affects my feelings
P2) does not touch me emotionally / touches me emotionally
P3) is not stimulating / is stimulating
P4) does not reach out to me / reaches out to me
P5) is not stirring / is stirring
P6) is not moving / is moving
P7) is not exciting / is exciting

Second, Lehto et al. (2012) developed a model with factors that predict perceived persuasiveness, and as part of this also considered the internal consistency of items to measure these factors. Several of their factors (e.g., dialogue support, design aesthetics) are not directly about persuasive messages *per se* but rather about the overarching behavioral intervention system they were studying. The aim of their work was not to develop a scale, so they did not try to develop factors that are independent of each other, but were mainly interested in how the factors related to each other. In fact, despite finding adequate internal consistency, they found quite a lot of cross-loadings, with items from one factor loading above 0.5 on other factors as well. Their validation was only in the health domain, and many of their questions specifically related to their intervention (e.g., a primary task support item "NIV provides me with a means to lose weight," a dialogue support item "NIV provides me with appropriate counseling," a perceived credibility item "NIV is made by health professionals"). So, this work did not result in a multi independent factors scale that can be used in multiple domains, like the scale developed in this paper. Considering the factors they considered, Perceived Credibility overlaps with the Quality factor in our scale (cf. trustworthy). Primary Task

support is related to the Effectiveness factor in our scale (e.g., "helps me change [my behavior]" is related to "causes a change in my behavior"). Their Perceived Persuasiveness factor has some relation to our Capability factor (e.g., compare "has an influence on me" and "has the potential to influence user behavior," "makes me reconsider [my behavior]," and "has the potential to change user behavior").

Third, Allen et al. (2000) compared the persuasiveness of statistical and narrative evidence in a message, and produced two scales to perform this study: a Credibility scale (measuring the extent to which one trusts the message writer) and an Attitude scale (measuring the extent to which one accepts the message's conclusion). They checked that each scale only contained one factor, and that each scale was internally consistent (in terms of Cronbach's alpha). They did not, however, consider whether items from one scale cross-loaded onto the other scale (e.g. the items "I think the writer is wrong" from the Attitude scale and "the writer is dishonest" from the Credibility scale seem related, so cross-loadings may well occur). They also did not remove an item with low factor loading ("the writing style is dynamic," loading 0.40) from the Credibility scale, which may indicate a poor scale structure (MacCallum et al., 1999). Their scales only measure some aspects of persuasiveness; for example, they do not measure the message's potential to inspire, or to cause behavior change.

Fourth, Popova et al. (2014), Jasek et al. (2015), and Yzer et al. (2015) used multi-item scales, but without a development phase. Popova et al. (2014) used five items (convincing-unconvincing, effective-ineffective, believable-unbelievable, realistic-unrealistic, and memorable-not memorable), Jasek et al. (2015) 13 (boring, confusing, convincing, difficult to watch, informative, made me want to quit smoking, made me want to smoke, made me stop and think, meaningful to me, memorable, powerful, ridiculous, terrible), and Yzer et al. (2015) 7 (convincing, believable, memorable, good, pleasant, positive, for someone like me). There is considerable overlap between these items and the ones we used for the scale development, though there are some items in these papers that seem more related to usability (e.g., "confusing") and some more related to feelings (e.g., "pleasant," "terrible").

Fifth, McLean et al. (2016) developed a scale from 13 items for measuring the persuasiveness of messages to reduce stigma about bulimia. They only performed an exploratory factor analysis (using ratings of only 10 messages), so no real validation. Their scale has two factors; one they describe as convincingness and the other as likelihood of changing attitudes toward bulimia. The first factor includes items such as "believable" and "convincing," which were part of our initial items for scale development and are related to the Quality factor in our scale. The second factor is related to the Capability factor of our scale.

In summary, the scale developed in this paper is unique in that it was developed from a large set of items covering a wide range of aspects of persuasiveness, was developed and validated across two domains, and has been shown to consist of three independent factors, with good internal consistency. The comparison of scale content with the content of other scales shows that the scale also provides reasonable coverage of concepts deemed important in the literature (for example, some aspects of Ethos, Pathos, and Logos are present).

5.2. Persuasiveness of Message Types

As a side effect of our studies, we also gained insights into the persuasiveness of message types. There have been several other papers investigating this, though these studies have only investigated the impact of Cialdini's principles and not the finer-grained argumentation schemes. For instance, Orji et al. (2015) and Thomas et al. (2017) investigated the persuasiveness of Cialdini's principles for healthy eating, Smith et al. (2016) for reminders to cancer patients, Ciocarlan et al. (2018) for encouraging small acts of kindness, and Oyibo et al. (2017) in general without mentioning specific domains.

Thomas et al. (2017) found that Authority messages were most persuasive and Liking least persuasive. Orji et al. (2015) found that Commitment and Reciprocity were the most persuasive over all ages and gender, whereas Consensus and Scarcity were the least persuasive. They found that females responded better to Reciprocity, Commitment, and Consensus messages than males. They also observed that adults responded better to Commitment than younger adults, and younger adults responded better to Scarcity than adults. Smith et al. (2016) observed that Authority and Liking were the most popular for the first reminder, and there was a preference for using Scarcity and Commitment for the second reminder. Ciocarlan et al. (2018) found that the Scarcity message worked best. Oyibo et al. (2017) observed that their participants were more susceptible to Authority, Consensus, and Liking.

The conflicting results of these studies can have several causes. Firstly, the studies were conducted in different domains. Our studies in this paper have shown that the persuasiveness of message types is in fact domain dependent. For example, we found in the Healthy Eating domain that some of the Authority-linked argumentation schemes scored badly on Effectiveness, and one of them was also worst on persuasiveness overall, whilst in the Email Security domain Authority-linked argumentation schemes scored best. Secondly, the studies used very different (and not validated) ways of measuring persuasiveness. So, it would be interesting to repeat all of these studies in a variety of domains using the scale developed in this paper. Thirdly, these studies did not consider the finer-grained argumentation schemes, but only Cialdini's principles. It is possible that, for example, the Authority messages used in one study followed a different argumentation scheme (within the Authority set) than those in another study. Finally, in contrast to our studies, none of these papers considered the individual factors of persuasiveness, but only considered persuasiveness as a whole. Our studies show that it is possible for a message type to score badly on one dimension on persuasiveness whilst scoring well on the others.

In summary, the most important results in this paper regarding the persuasiveness of message types are that (1) this persuasiveness is domain dependent, (2) investigating the finer-grained argumentation schemes matters as different results can be obtained for different argumentation schemes that are linked to the same Cialdini's principles, and (3) investigating the

different factor of persuasiveness matters as different results can be obtained for the different factors.

6. CONCLUSIONS

In this paper, we developed and validated a perceived persuasiveness scale to be used when conducting studies on digital behavior interventions. We conducted two studies in different domains to develop and validate this scale, namely in the healthy eating domain and the email security domain. The validated scale has 3 factors (Effectiveness, Quality, and Capability) and 9 scale items as illustrated in **Table 6**. We also discussed how this scale relates to and extends on earlier work on persuasiveness scales.

In addition to developing a scale, and to show its usefulness, we analyzed the impact of message types on the different developed scale factors. We found that message type significantly impacts on Effectiveness, Quality, and overall perceived persuasiveness in studies in both the healthy eating and email security domains. We also found a significant impact of message type on Capability in the email security domain. The three factors (as shown in the validation) measure different aspects of perceived persuasiveness. One example where this can also be seen is for the ARGUMENT FROM EXPERT OPINION WITH GOAL message type, which performs relatively badly on Effectiveness in the healthy eating domain but well on Quality in that domain. The persuasiveness of messages is clearly domain dependent. Additionally, our studies show that it is worthwhile to investigate the finer-grained argumentation schemes rather than just Cialdini's principles. We discussed related work on measuring the persuasiveness of message types and explained the conflicting findings in those studies.

As shown in our literature review, researchers working on digital behavior interventions tend to use their own scales, without proper validation of those scales, to investigate perceived persuasiveness. The validated scale developed in this paper can be used to improve such studies and will make it easier to compare the results of different studies and in different domains. We plan to use the scale to study the impact of message personalization across domains.

The work presented in this paper has several limitations. Firstly, we validated the scale in two domains (healthy eating and email security), and this validation needs to be extended to more domains. Secondly, the scale reliability needs to be verified. To investigate this, we need to perform a test-retest experiment in which participants complete the same scale on the same items twice, with an interval of several days between the two measurements. This also would need to be done in multiple domains. Thirdly, we need to repeat our studies into the impact message types with more messages and in more domains.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by CoPs ethics committee University of Aberdeen. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RT and JM contributed to the conception and design of the study. RT implemented the study, performed the statistical analysis, and wrote the first draft of the manuscript. All authors wrote sections of the manuscript, contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work on cyber-security in this manuscript was supported by the EPSRC under Grant EP/P011829/1.

REFERENCES

Allen, M., Bruflat, R., Fucilla, R., Kramer, M., McKellips, S., Ryan, D. J., et al. (2000). Testing the persuasiveness of evidence: combining narrative and statistical forms. *Commun. Res. Rep.* 17, 331–336. doi: 10.1080/08824090009388781

Anagnostopoulou, E., Magoutas, B., Bothos, E., Schrammel, J., Orji, R., and Mentzas, G. (2017). "Exploring the links between persuasion, personality and mobility types in personalized mobility applications," in *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*, eds P. W. de Vries, H. Oinas-Kukkonen, L. Siemons, N. Beerlage-de Jong, and L. van Gemert-Pijnen (Cham: Springer International Publishing), 107–118.

Busch, M., Patil, S., Regal, G., Hochleitner, C., and Tscheligi, M. (2016). "Persuasive information security: techniques to help employees protect organizational information security," in *Persuasive Technology*, eds A. Meschtscherjakov, B. De Ruyter, V. Fuchsberger, M. Murer, and M. Tscheligi (Cham: Springer International Publishing), 339–351.

Busch, M., Schrammel, J., and Tscheligi, M. (2013). *Personalized Persuasive Technology – Development and Validation of Scales for Measuring Persuadability*. Berlin; Heidelberg: Springer, 33–38.

Chang, J.-H., Zhu, Y.-Q., Wang, S.-H., and Li, Y.-J. (2018). Would you change your mind? an empirical study of social impact theory on facebook. *Telem. Inform.* 35, 282–292. doi: 10.1016/j.tele.2017.11.009

Cialdini, R. B. (2009). *Influence: The Psychology of Persuasion*. New York, NY: HarperCollins e-books.

Ciocarlan, A., Masthoff, J., and Oren, N. (2018). "Kindness is contagious: study into exploring engagement and adapting persuasive games for wellbeing," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18 (New York, NY: ACM), 311–319.

Cook, A., Pryer, J., and Shetty, P. (2000). The problem of accuracy in dietary surveys. Analysis of the over 65 UK national diet and nutrition survey. *J. Epidemiol. Commun. Health* 54, 611–616. doi: 10.1136/jech.54.8.611

Dziuban, C. D., and Shirkey, E. C. (1980). Sampling adequacy and the semantic differential. *Psychol. Rep.* 47, 351–357. doi: 10.2466/pr0.1980.47.2.351

Feltham, T. S. (1994). Assessing viewer judgement of advertisements and vehicles: scale development and validation. *ACR North Am. Adv.* 21, 531–535.

Grasso, F., Cawsey, A., and Jones, R. (2000). Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *Int. J. Hum. Comput. Stud.* 53, 1077–1115. doi: 10.1006/ijhc.2000.0429

Ham, C.-D., Nelson, M. R., and Das, S. (2015). How to measure persuasion knowledge. *Int. J. Advertis.* 34, 17–53. doi: 10.1080/02650487.2014.994730

Hammer, S., Lugrin, B., Bogomolov, S., Janowski, K., and André, E. (2016). "Investigating politeness strategies and their persuasiveness for a robotic elderly assistant," in *Persuasive Technology*, eds A. Meschtscherjakov, B. De Ruyter, V. Fuchsberger, M. Murer, and M. Tscheligi (Cham: Springer International Publishing), 315–326.

Hossain, M. T., and Saini, R. (2014). Suckers in the morning, skeptics in the evening: time-of-day effects on consumers' vigilance against manipulation. *Market. Lett.* 25, 109–121. doi: 10.1007/s11002-013-9247-0

Howitt, D., and Cramer, D. (2014). *Introduction to SPSS Statistics in Psychology*. Pearson Education.

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equat. Model. A Multidiscipl. J.* 6, 1–55. doi: 10.1080/10705519909540118

Jasek, J. P., Johns, M., Mbamalu, I., Auer, K., Kilgore, E. A., and Kansagra, S. M. (2015). One cigarette is one too many: evaluating a light smoker-targeted media campaign. *Tobacco Control* 24, 362–368. doi: 10.1136/tobaccocontrol-2013-051348

Kaptein, M., Markopoulos, P., de Ruyter, B., and Aarts, E. (2009). *Can You Be Persuaded? Individual Differences in Susceptibility to Persuasion*. Berlin; Heidelberg: Springer, 115–118.

Koch, T., and Zerbac, T. (2013). Helpful or harmful? How frequent repetition affects perceived statement credibility. *J. Commun.* 63, 993–1010. doi: 10.1111/jcom.12063

Lehto, T., Oinas-Kukkonen, H., and Drozd, F. (2012). "Factors affecting perceived persuasiveness of a behavior change support system," in *Thirty Third International Conference on Information Systems, Orlando*. Orlando.

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychol. Methods* 4:84.

MacKenzie, S. B., and Lutz, R. J. (1989). An empirical examination of the structural antecedents of attitude toward the ad in an advertising pretesting context. *J. Market.* 53, 48–65.

Mazzotta, I., de Rosis, F., and Carofiglio, V. (2007). Portia: a user-adapted persuasion system in the healthy-eating domain. *IEEE Intell. Syst.* 22, 42–51. doi: 10.1109/MIS.2007.115

McLean, S. A., Paxton, S. J., Massey, R., Hay, P. J., Mond, J. M., and Rodgers, B. (2016). Identifying persuasive public health messages to change community knowledge and attitudes about bulimia nervosa. *J. Health Commun.* 21, 178–187. doi: 10.1080/10810730.2015.1049309

Meschtscherjakov, A., Gärtner, M., Mirnig, A., Rödel, C., and Tscheligi, M. (2016). "The persuasive potential questionnaire (PPQ): Challenges, drawbacks, and lessons learned," in *Persuasive Technology*, eds A. Meschtscherjakov, B. De Ruyter, V. Fuchsberger, M. Murer, and M. Tscheligi (Cham: Springer International Publishing), 162–175.

Oduor, M., and Oinas-Kukkonen, H. (2017). "Commitment devices as behavior change support systems: a study of users' perceived competence and continuance intention," in *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*, eds P. W. de Vries, H. Oinas-Kukkonen, L. Siemons, N. Beerlage-de Jong, and L. van Gemert-Pijnen (Cham: Springer International Publishing), 201–213.

O'Keefe, D. J. (2018). Message pretesting using assessments of expected or perceived persuasiveness: evidence about diagnosticity of relative actual persuasiveness. *J. Commun.* 68, 120–142. doi: 10.1093/joc/jqx009

Orji, R. (2014). "Exploring the persuasiveness of behavior change support strategies and possible gender differences," in *Conference of 2nd International Workshop on Behavior Change Support Systems*, Vol. 1153, eds L. van Gemert-Pijnen, S. Kelders, A. Oorni, and H. Oinas-Kukkonen (Aachen: CEUR-WS), 41–57.

Orji, R., Mandryk, R. L., and Vassileva, J. (2015). "Gender, age, and responsiveness to cialdini's persuasion strategies," in *Persuasive Technology*, eds T. MacTavish and S. Basapur (Cham: Springer International Publishing), 147–159.

Orji, R., Vassileva, J., and Mandryk, R. L. (2014). Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. *User Model. User Adapt. Interact.* 24, 453–498. doi: 10.1007/s11257-014-9149-8

Oyibo, K., Orji, R., and Vassileva, J. (2017). "Investigation of the influence of personality traits on cialdini's persuasive strategies," in *Proceedings of the 2nd International Workshop on Personalization in Persuasive Technology*, Vol. 1833, eds R. Orji, M. Reisinger, M. Busch, A. Dijkstra, M. Kaptein, and E. Mattheiss (CEUR-WS), 8–20.

Popova, L., Neilands, T. B., and Ling, P. M. (2014). Testing messages to reduce smokers' openness to using novel smokeless tobacco products. *Tobacco Control* 23, 313–321. doi: 10.1136/tobaccocontrol-2012-050723

Purpura, S., Schw, V., Williams, K., Stubler, W., and Sengers, P. (2011). "Fit4life: The design of a persuasive technology promoting healthy behavior and ideal weight," in *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011* (Vancouver: ACM), 423–432.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res.* 99, 323–338. doi: 10.3200/JOER.99.6.323-338

Smith, K. A., Dennis, M., and Masthoff, J. (2016). "Personalizing reminders to personality for melanoma self-checking," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (Halifax), 85–93.

Thomas, R. J., Masthoff, J., and Oren, N. (2017). "Adapting healthy eating messages to personality," in *Persuasive Technology. 12th International Conference, PERSUASIVE 2017, Proceedings* (Amsterdam: Springer), 119–132.

Thomas, R. J., Oren, N., and Masthoff, J. (2018). "ArguMessage: a system for automation of message generation using argumentation schemes," in *Proceedings of AISB Annual Convention 2018, 18th Workshop on Computational Models of Natural Argument* (Liverpool), 27–31.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. New York, NY: Cambridge University Press.

Wells, S., Kotkanen, H., Schlafli, M., Gabrielli, S., Masthoff, J., Jylhå, A., et al. (2014). Towards an applied gamification model for tracking, managing, & encouraging sustainable travel behaviours. *EAI Endors. Trans. Ambient Syst.* 1:e2. doi: 10.4108/amsys.1.4.e2

Yzer, M., LoRusso, S., and Nagler, R. H. (2015). On the conceptual ambiguity surrounding perceived message effectiveness. *Health Commun.* 30, 125–134. doi: 10.1080/10410236.2014.974131

Zhang, K. Z., Zhao, S. J., Cheung, C. M., and Lee, M. K. (2014). Examining the influence of online reviews on consumers' decision-making: a heuristicâĂŞsystematic model. *Decis. Support Syst.* 67, 78–89. doi: 10.1016/j.dss.2014.08.005

Zhao, X., Strasser, A., Cappella, J. N., Lerman, C., and Fishbein, M. (2011). A measure of perceived argument strength: aeliability and validity. *Commun. Methods Meas.* 5, 48–75. doi: 10.1080/19312458.2010.547822

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer KS declared a past collaboration with one of the authors JM to the handling editor.

frontiers in Artificial Intelligence

OPINION
published: 28 April 2020
doi: 10.3389/frai.2020.00027

Digital Normativity: A Challenge for Human Subjectivation

*Eric Fourneret and Blaise Yvert**

Inserm and Univ Grenoble Alpes, BrainTech Lab U1205, Gières, France

Keywords: artificial intelligence, machine learning, free will (freedom), agency, ethics, education, normativity, governance

INTRODUCTION

Recent advances in artificial intelligence (AI) have opened unprecedented opportunities to humans to think and operate the world and its increasing complexity with digital technologies. Striking examples are deep neural networks (DNNs) (Lecun et al., 2015; Mnih et al., 2015), which can be trained quickly on large datasets either self-generated or already available from human experience. In particular, algorithms can become more efficient than humans on specific tasks after relatively short training periods compared to the time that humans need to learn (few hours or days, as compared to years). Their technical efficiency has for instance been demonstrated for optimizing financial transactions, speech or text recognition (Hinton et al., 2012), language translation (Hassan et al., 2018), real-time image content analysis, autonomous driving (Chen et al., 2015), or playing chess or go (Silver et al., 2017). They also start to see use in medicine to reach diagnoses (Lehman et al., 2019; Ye et al., 2019) and improve neuroprosthetics (Bocquelet et al., 2016; Schwemmer et al., 2018; Anumanchipalli et al., 2019). This multiplicity of technical demonstrations is thus progressively bringing AI central and ubiquitous in human life. Yet, the effectiveness of algorithms in bringing more and more relevant recommendations to humans may start to compete with human-alone decisions based on values other than pure efficacy. Here, we examine this tension in light of the emergence of several forms of digital normativity, and analyze how this normative role of AI may influence the ability of humans to remain subject of their life.

THE ADVENT OF DIGITAL NORMATIVITY

The increasing role of AI is engendering the emergence of several forms of digital normativity, the ability of algorithms to establish standards that humans incorporate as what should be considered as normal in their lives and guide their actions. First, algorithms tend to reproduce the trends that are most present in the data on which they have been trained. This creates a normalized view of the problem they are intended to solve. The level of details that algorithms might be able to discriminate can be high, as for instance, in automated image pattern recognition or autonomous driving (Kaur and Rampersad, 2018). This first form of digital normativity may thus often be satisfying enough for humans to rely on algorithmic recommendations. However, the automatic and thus objective processing of large datasets restitutes general trends present in these datasets, whether ethically good or bad (Hardt et al., 2016).

Another form of digital normativity arises from the use of predictive algorithms trained on objective observational data without accounting for the course through which this data has been generated. For instance, algorithms that provide a customer with purchasing suggestions only rely on previous purchases made by the same and other customers, without access to the personal reasons underlying these purchases. This form of automatic data processing thus eliminates the inherent subjectivity of the customer: The individual is objectivized (normalized) by the algorithm

OPEN ACCESS

Edited by:
Fridolin Wild,
Oxford Brookes University,
United Kingdom

Reviewed by:
Rebecca Raper,
Oxford Brookes University,
United Kingdom
Matthias Rolf,
Oxford Brookes University,
United Kingdom, in collaboration with reviewer RR
Julita Vassileva,
University of Saskatchewan, Canada

***Correspondence:**
Blaise Yvert
blaise.yvert@inserm.fr

Specialty section:
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

Received: *14 October 2019*
Accepted: *31 March 2020*
Published: *28 April 2020*

Citation:
Fourneret E and Yvert B (2020) Digital Normativity: A Challenge for Human Subjectivation. Front. Artif. Intell. 3:27. doi: 10.3389/frai.2020.00027

(Ayres, 2007). This second form of digital normativity is actually a recursive and dynamic process: Algorithmic recommendations emanating from previous human actions in turn influence their next actions (Rouvroy and Berns, 2013; Thomassey and Zeng, 2018).

The normative role of algorithms takes a third form when their efficiency outperforms that of humans. If, for a given application, an algorithm has a higher predictive power than any human expert, it may indeed become reasonable to rely solely on this algorithm to make decisions. The algorithm then creates the norm by imposing its efficacy. The efficiency becoming the norm, the question becomes whether the role of humans in determining for themselves the finality of this efficiency could be challenged.

DIGITAL NORMATIVITY AND SUBJECTIVATION

Subjectivation (Wieviorka, 2012) is a construction process leading someone to become and be aware of being a subject, i.e., being free and responsible for one's actions and at the foundation of one's representations and judgments. This capacity is progressively acquired throughout life experience, including education, professional life, and more. Given that AI now constitutes an important part of human environment, could this technology weaken, or on the contrary, help to boost the capacity of human individuals to become subjects of their individual and collective lives?

Such question could be considered as irrelevant since it is humans who develop AI algorithms. This role ensures that the human action remains required, and if algorithms help to make decisions, their recommendations still result from a set of rules established by humans. However, AI algorithms may still influence the process of subjectivation. For instance, a search engine giving access to a huge amount of available knowledge in just a few clicks offers unique opportunities to any individual to build his or her critical judgment, and thus to become a human subject. The same engine may also bias subjectivation when results put at the top of the list are based on a statistical inference that does not account for the user as a subject.

Once subjectivation has been acquired, AI may further influence how it is exerted. Humans may indeed no longer desire to make decisions by themselves whenever algorithms may efficiently handle for them this task. This could be for the sake of either physical comfort when an action is physically demanding (e.g., driving long distances), or psychological comfort when a decision engages a moral responsibility difficult to endorse. As such, algorithms are used in the justice system in Belgium to evaluate the risk of recidivism and help determine whether an imprisoned individual should benefit from anticipatory freedom. In this scenario, the judges' responsibility may be increased and more difficult to stand if they decide against the recommendation of an algorithm. If their decision is indeed found later to be inappropriate, they could be opposed to have acted against an algorithmic decision considered more objective than a human decision (Rouvroy and Berns, 2013). Although it remains theoretically possible to resist such normativity, the associated amplification of human responsibility could become so much of a deterrent that disobedience would become difficult or even no longer possible in practice. An increasing number of opportunities may therefore be offered to humans to progressively disengage from their role of subjects of their lives (Erel et al., 2019), leading to the emergence of certain forms of governance without subject.

THE RISK OF A SILENT HUMAN DESUBJECTIVATION

Despite their importance in the organization of human societies, algorithms do not decide alone and a cooperative relationship between humans and AI exists: on the one hand a form of expertise (the algorithm) and on the other hand the power to decide (humans). Each needs the other but both do not merge as one. Indeed, a competence to make decisions differs from a competence of expertise: A power to decide can be exerted in absence of expertise, and conversely, an expert is not necessarily competent to decide (Green, 2012; Heitz, 2013). Deciding is acting with doubts, thus accepting the risk of making errors. If humans were to refuse this risk and transfer their power of decision to more efficient algorithms, they would jeopardize an essential part of their humanity: their ability to learn from errors and thus their power of perfectibility (Rousseau, 1754).

Current generations remain vigilant regarding this risk but what about future generations born after the emergence of digital normativity, and thus well-habituated with its ubiquity? When introducing the notion of voluntary servitude, La Boétie already seized this question to understand the foundations of despotic political powers. He pointed out that in a process of oppression, people are at first aware of losing their freedom but that the next generations make this situation of oppression the rule and become unaware of their servitude or accustomed to it: "(...) Those who come after serve without regret, and willingly do what their predecessors had done by constraint" (La Boétie, 1576). Importantly, the advent of AI governmentality would not impose itself by any violent physical or moral means, but by meddling into human life through progressive changes of practice. This is where a risk of silent human desubjectivation could take root.

This risk is further strengthened by the challenge of explicability of AI algorithms. Although the methods used to train algorithms are well-understood (e.g., backpropagation), the resulting set of optimal parameters does not generally represent any intuitive or ecological meaning for a human being. Then the question is: Can we ethically follow a recommendation deduced through a reasoning surpassing human expertise but no longer accessible? The risk would be to make decisions blindly without critical evaluation, thus silencing the capability of the human subject to distinguish between the fair and the unfair.

CONCLUSION: THE NECESSITY OF AN ETHICS BY DESIGN

AI has clearly become a unique opportunity for accompanying the evolution of human well-being but engenders a new major ethical challenge for humans: to preserve our capability to remain subjects and not only agents. Far from either completely

embracing or completely rejecting AI technologies, it has become essential that an ethical reflection accompany the current developments of intelligent algorithms beyond the sole question of their social acceptability. Such thoughtful reflection cannot be conducted independently from the scientific actors of AI technology, but needs to accompany them in defining the values and aims of their research. The Ethics-by-design methodology introduced by Verbeek (2011) can be used for such purpose. When designing a new technology, this methodology consists first in identifying the system of values of the technology (e.g., the power of objective prediction of AI and its efficiency in extracting relevant features of massive amount of data), and then in thinking the principles of protection of the subjectivation process from the beginning of the conception of the technology (e.g., how a speech neural prosthesis can be conceived in such a way that the externalization of the user's inner speech remains under his full control, Rainey et al., 2018). In practice, ethics by design can be implemented by anchoring philosophers and ethicists in scientific groups developing the technologies. Moreover, educational programs toward the next generations of scientists born with AI and dedicated to the ethical implications of AI would also be key elements to ensure the perenity of such ethical reflection. This double scientific and societal anchoring of a pragmatic ethics is mandatory to preserve human subjectivation, free will, and freedom in the long term: "Techniques always bring with them the world in which they will make sense" (Guchet, 2014). AI should not be developed to invent the future for us, but to help us invent our future.

AUTHOR CONTRIBUTIONS

EF and BY conducted this reflection and wrote the manuscript.

FUNDING

This work was supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 732032 (BrainCom), and by the French National Research Agency under Grant Agreement No. ANR-16-CE19-0005-01 (Brainspeak).

REFERENCES

Anumanchipalli, G. K., Chartier, J., and Chang, E. F. (2019). Intelligible speech synthesis from neural decoding of spoken sentences. *Nature* 568, 493–498. doi: 10.1038/s41586-019-1119-1

Ayres, I. (2007). *Super Crunchers. Why Thinking-by-Numbers is the New Way to be Smart*. Bantam; Reprint Edition.

Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., and Yvert, B. (2016). Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLOS Comput. Biol.* 12:e1005119. doi: 10.1371/journal.pcbi.1005119

Chen, C., Seff, A., Kornhauser, A., and Xiao, J. (2015). "DeepDriving: learning affordance for direct perception in autonomous driving," in *Proceeding IEEE ICCV*, 2722–2730. doi: 10.1109/ICCV.2015.312

Erel, I., Stern, L. H., Tan, C., Weisbach, M. S., and Selecting Directors Using Machine Learning (2019). *Fisher College of Business Working Paper No. 2018-03-005*, Finance Working Paper No. 605/2019. European Corporate Governance Institute (ECGI). Available online at: https://ssrn.com/abstract=3144080

Green, C. (2012). Nursing intuition: a valid form of knowledge. *Nurs. Philos.* 13, 98–111. doi: 10.1111/j.1466-769X.2011.00507.x

Guchet, X. (2014). *Philosophie des Nanotechnologies*. HR.EVOL.MA., ed. Paris:Hermann Paris.

Hardt, M., Price, E., and Srebro, N. (2016). "Equality of opportunity in supervised learning," in *Proceeding NIPS*, 3323–3331.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., et al.. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv[Preprint].arXiv:1803.05567*.

Heitz, J. (2013). La décision : ses fondements et ses manifestations. *RIMHE* 1, 106–117. doi: 10.3917/rimhe.005.0106

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* 29, 82–97. doi: 10.1109/MSP.2012.2205597

Kaur, K., and Rampersad, G. (2018). Trust in driverless cars: investigating key factors influencing the adoption of driverless cars. *J. Eng. Tech. Manag.* 48, 87–96. doi: 10.1016/j.jengtecman.2018.04.006

La Boétie, E. (1576). *Discours de la servitude volontaire ou le Contr'un*.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lehman C. D., Yala, A., Schuster, T., Dontchos, B, Bahl, M., Swanson, K., et al. (2019). Mammographic breast density assessment using deep learning: Clinical implementation. *Radiology* 290, 52–58. doi: 10.1148/radiol.2018180694

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Rainey, S., Maslen, H., Megevand, P., Arnal, L., Fourneret, E., and Yvert, B. (2018). Neuroprosthetic speech: the ethical significance of accuracy, control and pragmatics. *Cambridge Q. Healthc. Ethics*. 28, 657–670. doi: 10.1017/S0963180119000604

Rousseau, J. (1754). *Discours sur l'origine et les fondements de l'inégalité parmi les hommes*. Œuvres Complètes.

Rouvroy, A., and Berns, T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux* 177, 163–196. doi: 10.3917/res.177.0163

Schwemmer, M. A., Skomrock, N. D., Sederberg, P. B., Ting, J. E., Sharma, G., Bockbrader, M. A., et al. (2018). Meeting brain–computer interface user performance expectations using a deep neural network decoding framework. *Nat. Med.* 24, 1669–1676. doi: 10.1038/s41591-018-0171-y

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature* 550, 354–359. doi: 10.1038/nature24270

Thomassey, S., and Zeng, X. (eds.). (2018). "Artificial intelligence for fashion industry," in *The Big Data Era,* Thomassey (Singapore: Springer) doi: 10.1007/978-981-13-0080-6

Verbeek, P.-P. (2011). *Moralizing Technology*. Chicago: The University Chicago Press. doi: 10.7208/chicago/9780226852904.001.0001

Wieviorka, M. (2012). *Du concept de sujet à celui de subjectivation / dé-subjectivation*. FMSH-WP-2012–P-2016.

Ye, W., Gu, W., Guo, X., Yi, P., Meng, Y., Han, F., et al. (2019). Detection of pulmonary ground-glass opacity based on deep learning computer arti fi cial intelligence. *Biomed. Eng. Online* 18, 1–12. doi: 10.1186/s12938-019-0627-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 20 January 2021
doi: 10.3389/frai.2020.595627

Automated Disengagement Tracking Within an Intelligent Tutoring System

Su Chen[1,2]*, *Ying Fang*[2,3], *Genghu Shi*[2,3], *John Sabatini*[2,3], *Daphne Greenberg*[4], *Jan Frijters*[5] *and Arthur C. Graesser*[2,3]

[1]Department of Mathematical Sciences, University of Memphis, Memphis, TN, United States, [2]Institute for Intelligent Systems, University of Memphis, Memphis, TN, United States, [3]Department of Psychology, University of Memphis, Memphis, TN, United States, [4]Department of Learning Sciences, Georgia State University, Atlanta, GA, United States, [5]Department of Child and Youth Studies, Brock University, St. Catharines, ON, Canada

This paper describes a new automated disengagement tracking system (DTS) that detects learners' maladaptive behaviors, e.g. mind-wandering and impetuous responding, in an intelligent tutoring system (ITS), called AutoTutor. AutoTutor is a conversation-based intelligent tutoring system designed to help adult literacy learners improve their reading comprehension skills. Learners interact with two computer agents in natural language in 30 lessons focusing on word knowledge, sentence processing, text comprehension, and digital literacy. Each lesson has one to three dozen questions to assess and enhance learning. DTS automatically retrieves and aggregates a learner's response accuracies and time on the first three to five questions in a lesson, as a baseline performance for the lesson when they are presumably engaged, and then detects disengagement by observing if the learner's following performance significantly deviates from the baseline. DTS is computed with an unsupervised learning method and thus does not rely on any self-reports of disengagement. We analyzed the response time and accuracy of 252 adult literacy learners who completed lessons in AutoTutor. Our results show that items that the detector identified as the learner being disengaged had a performance accuracy of 18.5%, in contrast to 71.8% for engaged items. Moreover, the three post-test reading comprehension scores from Woodcock Johnson III, RISE, and RAPID had a significant association with the accuracy of engaged items, but not disengaged items.

Keywords: intelligent tutoring system, conversational agents, AutoTutor, mind wandering, disengagement

OPEN ACCESS

Edited by:
H. Chad Lane, University of Illinois at Urbana-Champaign, United States

Reviewed by:
Ranilson Oscar Araújo Paiva, Federal University of Alagoas, Brazil
Jason Bernard, University of Saskatchewan, Canada

***Correspondence:**
Su Chen
schen4@memphis.edu

Specialty section:
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

Received: *17 August 2020*
Accepted: *08 December 2020*
Published: *20 January 2021*

Citation:
Chen S, Fang Y, Shi G, Sabatini J, Greenberg D, Frijters J and Graesser AC (2021) Automated Disengagement Tracking Within an Intelligent Tutoring System. Front. Artif. Intell. 3:595627. doi: 10.3389/frai.2020.595627

INTRODUCTION

Many intelligent tutoring systems (ITSs) implement natural language dialogue and provide one-on-one human-like tutoring in an automated fashion (Woolf, 2010; Graesser et al., 2012; Nye et al., 2014; Graesser, 2016; Johnson and Lester, 2016; Graesser et al., 2017). A well-designed ITS offers personalized and adaptive instruction which is difficult (or sometimes impossible) to implement in a traditional classroom setting with a teacher handling 30 or more students. Some ITSs have been designed to be similar to human tutors in the design of content coverage and tutorial interaction patterns, such as AutoTutor and other systems with conversational agents that have similar architectures to AutoTutor (Nye et al., 2014; Graesser, 2016). Of particular relevance to the present study, ITS designers, human tutors, as well as classroom teachers struggle with how they can best keep the students focused and engaged in content learning. It is well established that engagement is an important component of learning and motivation (Csikszentmihalyi, 1990;

D'Mello and Graesser, 2012; Larson and Richards, 1991; Mann and Robinson, 2009; Pekrun et al., 2010; Pekrun and Stephens, 2012). An automated disengagement detector would be of benefit to students, as well as to tutors, teachers, and ITS environments.

Regardless of whether students learn from an ITS, a human tutor, or a teacher in a classroom, students are likely to become disengaged due to various reasons, such as fatigue, environmental distractions, loss of interest, or the stress of falling behind in a course, as will be elaborated below. One strategy that ITS developers have taken has been to increase engagement through gamification (Jackson and McNamara, 2013; Millis et al., 2017), but students can experience disengagement in games just as they do in learning environments without gamification. A different strategy is to detect disengagement as it occurs, so as to better intervene with the student, for example, by redirecting their attention to learning. The prediction and tracking of disengagement can be approached in different ways, such as developing models of disengagement (sometimes operationalized as boredom) from individual difference measures, language, and keystroke analyses (D'Mello and Graesser, 2012; Bixler and D'Mello, 2013; Allen et al., 2016). Tracking students' disengagement promptly would allow personalized interactions at appropriate times in order to re-engage students. A small number of studies have been conducted with personalized interventions to prevent or interrupt disengaging behaviors and guide an individual learner back on track (Woolf et al., 2010; D'Mello and Graesser, 2012; D'Mello et al., 2012; Lane, 2015; Bosch et al., 2016; Monkaresi et al., 2017). A critical component of such interventions is a built-in disengagement tracking algorithm which can capture behavioral disengagement promptly and accurately.

Disengagement occurs in a number of situations, such as when the student is 1) mind wandering (Feng et al., 2013; Smallwood and Schooler, 2015), 2) distracted by an extraneous goal, 3) impetuously responding in order to finish the task quickly without concern for performance, or 4) "gaming" the learning environment, such as having an adaptive system filling in most of the answers and solutions to problems (Baker et al., 2008). Multiple factors can lead to disengagement or "off-track" behaviors, and these can be voluntary or involuntary. The time-course of completing a task is also an important consideration. For example, students might begin a learning session in an ITS with some level of interest and enthusiasm, but boredom or fatigue may creep in as the session progresses, as the novelty of the system fades, or when they have difficulty comprehending as the material becomes progressively more complex. The latter is of particularly relevance to this study, as disengagement is negatively related to reading comprehension (Millis et al., 2017).

Disengagement also presents a problem for researchers interested in evaluating learning and performance. Time on task alone (e.g., time spent on one question, problem, text, or session) can be considered contaminated by disengagement in contrast to diligent efforts to complete the task. Disengaged students may take too long a time (thinking about something irrelevant to the reading task) or too short a time (quickly finishing the question or session without comprehension) on a given question, problem, text or session. That is, a disengaged reader can be extremely slow or fast in processing during a learning task with low performance. Data analyses that do not consider the abnormal reading time due to disengagement may lead to unreliable or even misleading results. Moreover, a simple unidimensional measure of time is not sufficiently diagnostic of disengagement because both very fast times and very slow times can be signals of disengagement.

Existing disengagement/engagement detection methods that focus on mind wandering have applied supervised learning approaches to train models using self-reported mind-wandering (Mills and D'Mello, 2015; Millis et al., 2017; Bosch and Dmello, 2019) or use of commercial eye-tracker to automatically detect mind-wandering (D'Mello et al., 2012; Hutt et al., 2019). Another approach uses researcher-defined disengagement when examining student performance profiles over days or weeks, such as a student who is inactive for at least seven consecutive days (Chen and Kizilcec, 2020). In the self-reported approach, the participants are probed during reading with a stimulus signal, upon which they report whether or not they are mind-wandering. Self-reported mind-wandering is not considered a practical tracking system for detecting concurrent disengagement, however, because such self-reports could interfere with the learning process. Moreover, these self-reports may have a response bias to the extent that disengaged students may not admit that they have been disengaged due to social desirability bias (Holden and Passey, 2010). An alternative to self-report was proposed by Beck (2005). In this approach, item response theory was used to predict the probability of a correct response based on the response time and then estimate the probability of disengagement given the probability of being correct for engaged vs. disengaged students. However, Beck's method requires a large sample size to build a model that accounts for inter-student and question variability since a large number of parameters were introduced. This method is therefore also not suitable for tracking disengagement during tutoring since the sample size required is only attained after a student completes a large number of questions. Additionally, existing methods mainly focus on detecting students that are disengaged rather than a specific period where a student gets disengaged (Bulathwela et al., 2020). It would be more helpful if we can detect the time period where students start to get disengaged and re-engage them promptly.

Graesser, Geenberg, Frijters, and Talwar (submitted) identified questions that a student answers that are within the student's "zone of engagement". These "engaged question-answer observations" included questions that were answered neither too slowly nor too quickly (within ± 0.5 standard deviation of mean log of response time), based on a student's personal average speed of answering questions in a lesson. The participants were struggling adult readers (N = 52) who completed up to 30 lessons in a computerized learning environment (AutoTutor) that was part of a 4-month intervention that trained them on comprehension strategies. Answer time alone was not sufficient to identify the incidence of disengagement because accuracy in answering the questions is obviously important. Therefore,

questions outside of the zone of engagement, "disengaged question-answer observations" were defined as being answered incorrectly and too quickly or slowly. This approach to identifying disengaged observations was completed after the 4-month study was completed. Unfortunately, this method, however, is not suitable for monitoring concurrent disengagement since disengaged question-answer observations can only be detected at the end of a reasonably large sample of lessons. That does not allow an intelligent learning environment to give feedback and guidance to the learner when disengagement is detected. Moreover, if a question is incorrectly and slowly answered, it may not necessarily indicate disengagement. It is possible that a student is at the very early stage of learning new material and spending time in productive comprehension activities. Nevertheless, an approach to detecting disengagement based on the accuracy and time to answer questions during training is a reasonable approach to building a disengagement tracking system. It does not require special physiological or neuropsychological sensing devices, eye tracking, self-reports of engagement, or machine learning with supervised training that cannot scale up to real-world applications. The approach would be more useful to the extent it could detect disengagement in a smaller time span, such as a minute or two.

In this paper, we propose an unsupervised self-learning algorithm to monitor whether a student is engaged in answering questions within lessons of a conversation-based intelligent tutoring system. The system is *AutoTutor for Adult Reading Comprehension* (AutoTutor-ARC), a version of AutoTutor to teach adult learners reading comprehension strategies. In AutoTutor systems, a tutor agent and optionally a peer agent hold conversations with a human student. When the conversation has two agents (tutor and peer), the conversations are called *trialogues*, as opposed to tutor-student dialogues (Millis et al., 2011; Graesser et al., 2014). Similar three way interactions between two agents and humans have been designed in other learning and assessment environments (Danielle et al., 2006; Jackson and McNamara, 2013; Zapata-Rivera et al., 2015; Lippert et al., 2020) and even in museums (Swartout et al., 2010). Disengagement is detected in an algorithm that considers the time that an adult student spends answering a question, and his/her performance accuracy (i.e. whether a question was answered correctly). Disengaged learners tend to spend too long or too short a time on a particular question and perform poorly on the question or adjacent questions (Greenberg et al.; Millis et al., 2017).

The proposed algorithm starts out by considering the first three to five correctly answered questions to estimate the students' engagement pace within a specific lesson. The underlying assumption is that students are engaged at the beginning phase of a lesson and most likely performing well. Engagement time to answer a question can be estimated at this early phase of a lesson and serve as a standard of engagement for a particular student on a particular lesson. Based on the standard, the algorithm subsequently tracks students' performance to identify questions for which they exhibit disengagement by virtue of being inaccurate or too fast or slow compared with the engagement pace. The underlying assumption is that students are engaged at the beginning phase of a lesson but periodically become disengaged in latter phases when they are bored, confused with difficult material (e.g. sometimes due to the increment in levels of difficulty designed in AutoTutor), or mind wandering. We implemented the proposed algorithm to predict/monitor disengagement in AutoTutor-ARC. Our results show that items that the detector identified as the learner being disengaged had a performance accuracy of 18.5%, in contrast to 71.8% for engaged items. Moreover, three post-test reading comprehension scores from Woodcock Johnson III, RISE, and RAPID had a significant association with the accuracy of engaged items, but not disengaged items. The development of DTS algorithm is motivated by response time and performance data generated by the users of AutoTutor-ARC system. DTS has not been used in any intelligent system yet. The validation analyses in the manuscript can be considered as a "low stakes" application of DTS. If successful at detecting disengagement, the proposed real-time disengagement tracking system could be of value in enhancing learning efficiency in future AutoTutor-ARC systems, if it can be coupled with interventions during a lesson that re-engage a disengaged student. The algorithm could also be applied to other computer-based learning or assessments that utilize a question-answer environment.

DATA

Description of AutoTutor-ARC

There are many versions of AutoTutor on various topics, strategies and skills that help students learn by holding a conversation in natural language with computer agents (Nye et al., 2014; Graesser, 2016). AutoTutor-ARC was developed to help adult learners improve reading comprehension. It was first implemented as part of an intervention study conducted by the Center for the Study of Adult Literacy (CSAL, http://csal.gsu.edu). AutoTutor-ARC is a web-based intelligent tutoring system with 30 lessons focusing on building reading comprehension strategies (Graesser et al., 2016b). In each lesson, the learner engages in tutored instruction on comprehension strategies by having trialogue conversations with two computer agents (a tutor and peer). Through the three-way conversations, the learners are provided not only with instructions on reading comprehension strategies, but also guided and hopefully motivated by the computer agents during the learning process.

The lessons typically start with a 2–3 min video that reviews the comprehension strategy that is the target of the lesson. After the review, the computer agents scaffold students through the learning by asking questions, providing short feedback, explaining how the answers are right or wrong, and filling in information gaps. Since adult learners in AutoTutor typically have substantial challenges in writing, AutoTutor tends to rely on point-and-click (or touch) interactions, multiple-choice questions, drag-and-drop functions, and other conventional input channels. The learner chooses the answer by selecting an answer, while the peer agent sometimes gives his answer by talking. Flow within each lesson is driven by either a fixed

sequence or contingent branching. The first set of question-answer items within a particular lesson is the same for all students who take the lesson. Fixed sequence lessons deliver the same set of conversational questions to all students, independent of their performance throughout the entire lesson. Contingent branching lessons start out with questions and materials at a medium level of difficulty, but subsequently shift to harder or easier materials/questions depending on their performance on the medium difficulty material (Graesser et al., 2017). For example, 11 of the lessons have multi-sentence texts. For each of these multi-sentence texts, students read the text and then are asked approximately 10 agent-based conversational questions in a fixed sequence for the text. If a student performs well on the 10 questions, then the student receives a second more difficult text with a fixed sequence of approximately 10 questions; students below mastery threshold receive a relatively easier text with approximately 10 questions. These questions are consecutively ordered within one lesson that a student receives. For example, if the first text has 10 questions, coded 1 to 10, and then the second text's questions start with 11 and go to 20. Thus, a lesson may contain questions of two different difficulty levels, e.g. "medium and easy" or "medium and hard". Some lessons have contingent branching but there is a smaller span of text to be comprehended, such as the comprehension of sentences or words in a sentence. Again, these question items start out medium but branch to more easy or difficult items depending on the student's performance. Accuracy (correct/incorrect) and time spent (called *response time* (RT) later in the manuscript) on each question is recorded per lesson per student.

Participants and Design

The data sets used to test the proposed algorithm were taken from three waves of an intervention study in two medium sized cities. Participants were 252 adult learners who were offered approximately 100 h of instructional intervention designed to improve their reading skills. The intervention period lasted over 4 months and was implemented in hybrid classes, which consisted of teacher-led sessions and the computer-based AutoTutor-ARC sessions. Their ages ranged from 16 to 74 years (M = 42.4, SD = 13.9) and 74.6% were female. The reading level of participants ranged from 3.0 to 7.9 grade equivalencies. On average, the 252 participants completed 30 lessons. The adult students were also assessed with three standardized tests of comprehension before and after the instruction.

The AutoTutor-ARC content (i.e., lessons and texts) were scaled according to Graesser and McNamara's (2011) multilevel theoretical framework of comprehension. The framework specifies six theoretical levels: word (W), syntax (Syn), the explicit textbase (TB), the referential situation model (RSM), the genre/rhetorical structure (RS), and the pragmatic communication. Words and syntax represent lower level basic reading components that include morphology, word decoding, word order and vocabulary (Rayner et al., 2001; Perfetti, 2007). The TB level focuses on the meaning of explicit ideas in the text, but not necessarily the precise wording and syntax. The RSM level refers to the subject matter and requires inferences to be made on the explicit text and it differs by text type. For example, in narrative text, the RSM includes the characters, objects, settings, events and other details of the story; while in informational text, the model corresponds to substantive subject matter such as topics and domain knowledge. Rhetorical structure/discourse genre (RS) focus on the differentiated functional organization of paragraphs and type of discourse, such as narration, exposition, persuasion and description. Among the four theoretical levels, TB, RSM and RS are assumed to be more advanced and difficult to master compared to words and syntax (Perfetti, 2007; Cain, 2010)). AutoTutor taps all of these levels except for syntax and pragmatic communication. Each lesson was assigned a measure of the relevance to one to three of the four theoretical levels according to the extent to which the level was targeted in this lesson. The assigned codes were primary, secondary, tertiary or no relevance of a component to a lesson, corresponding to a relevance score of 1.00, 0.67, 0.33 and 0.00 respectively (Shi et al., 2018). In this study, we simply consider the primary theoretical level that characterizes the lesson. **Table 1** specifies the primary theoretical levels that characterize the 34 lessons (Actually 34 lessons were designed in CSAL, but only 30 (or less) lessons were assigned to the 252 learners in pilot studies).

METHODOLOGY

Algorithm of Disengagement Tracking System

An automated disengagement tracking system (DTS) is ideally personalized to the response times of individual students who work on a particular lesson. For any given student, a DTS should adapt to the learner's pattern of engaged performance, that is, the typical response time when engaged in attending to lesson content. Disengagement is detected when a student's performance (reading time or accuracy in answering questions) significantly deviates from this 'typical' pattern. In AutoTutor-ARC, a student is asked to read a text or sentence and to answer questions that are woven into the conversation between the two agents and the student. The system records the time that this student spends on each question and whether a question is answered correctly (1: correct, 0: incorrect). The amount of time a student takes to respond to a question, namely the response time (RT), is one behavior that can be used to determine whether a student is disengaged while working on this question. Performance suffers when the student is disengaged. One indication that students are disengaged is that they respond too fast or too slow (relative to their personalized typical RT) on a question. Too short or long RT does not necessarily mean "disengagement", since other factors influence RT. However, the short and long times can often be signals that probabilistically predict disengagement. For example, a short RT could be impetuous responding or gaming the system, whereas a long RT may be a difficulty level shift in texts/questions, mind wandering, or a personal bio break. To supplement the validity of RT indicator, we can consider another indication of

TABLE 1 | Distribution of Primary Theoretical Levels Across the 34 lessons.

Theoretical level	Number of lessons	Lesson names
Word (W)[a]	4	4-Word Parts, 6-Word Meaning Clues, 7-Learning New Words, 8-Multiple Meaning Words
Textbase (TB)	4	9-Pronouns, 12-Key Information, 16-Main Ideas, 17-Persuasive Texts
Referential Situation Model (RSM)	15	1-Text Signals, 10-Non-Literal Language, 11-Review 1, 13-A Personal Story,14-Connecting Ideas, 15-Story Maps,18-Review 2, 27-Complex Stories,28-Inferences from Texts, 29-Complex Persuasive Texts, 30-Forms and Documents, 31-Job Applications, 32-Searching the Web, 33-Using Email, 34-Social Media
Rhetorical Structure (RS)	11	2-Purpose of Texts, 3-Complex Texts, 5-Punctuation, 19-Claims vs. Support, 20-Problems and Solutions, 21-Cause and Effect, 22-Describing Things, 23-Compare and Contrast, 24-Time and Order, 25-Steps in Procedures, 26-Review 3

[a]*Syntax is grouped into the words (W) category in* ***Table 1****.*

disengagement: a significant drop in the correctness rate of a student. Thus, the DTS detects questions that a student is disengaged using both indications together. A flow chart of the DTS algorithm is shown in **Figure 1**.

The top half of **Figure 1** demonstrates the process of identifying parameters of response time distribution of engaged question-answer observations (i.e. learning stage in **Figure 1**), while the bottom half provides a logical procedure for disengagement detection using the parameters estimated in the learning stage (i.e. detection stage in **Figure 1**). Evidence shows that engagement wanes as time passes and disengagement usually occurs in the later phase when a subject withdraws from the commitment to task goals (Millis et al., 2011; Hockey, 2013). Hence, it is reasonable to make an assumption that a learner is more likely to be engaged at the beginning of a lesson. DTS learns a student's engaged RT from the first few questions within a lesson and uses it to identify questions with abnormal (or disengaged) RT later on.

In the first phase at the beginning of a lesson, the DTS obtains the distribution of a student's engaged RT on questions. Response time is usually right-skewed, as is the current data set, so a log transformation was applied to make the data resemble a normal distribution. We assume an engaged student's $log(RT)$ on a question within a specific lesson is normally distributed with mean μ and standard deviation σ. In practice, most per person and per lesson $log(RT)$ distributions meet the normality assumption, or very close to it. This assumption was checked and validated before we started our analysis. To this end, we make two assumptions: 1) students tend to be engaged at the beginning of a lesson when answering the first few questions, and 2) if a student correctly answered a question, he/she is likely to be engaged. It is possible that students may be disengaged at the beginning of a lesson due to a variety of reasons. Alternatively, students may correctly answer a question by chance when they are actually disengaged. However, there is a low probability that a learner is disengaged and correctly answers several questions by guessing or randomly clicking. The proposed method focuses on the first few (e.g., five) questions that were correctly answered and assumes that the students were engaged while working on these questions that were correctly answered. Even though there might be very few questions that were mistakenly counted as "engaged" (when they should be counted as "disengaged"), the results of the proposed method should not be substantially affected since we excluded the extreme (minimum and maximum) RT of the initial questions, as will be elaborated below.

We will now turn to some of the mathematical specification of the DTS algorithm. Suppose that students are engaged on the first b correctly answered questions and start to get disengaged at the s^{th} question some time point later. Presumably, s should be greater than or equal to b for the system to learn a user's engaged response time in a specific lesson. If $s \leq b$, the algorithm will specify that you will need more questions to detect disengagement. If $s > b$, DTS will automatically treat the response time of the first b correctly answered questions as engaged response time. If there are less than b correct question-answer observations up to the s^{th} question, we tentatively use question #2 to question #b's response time as engaged response time instead. We excluded question #1 since the users usually take extra time to read the text in the first question and spend much longer time than usual. Let I be the first b correctly answered questions, whereas μ and σ are estimated by

$$\widehat{\mu} = \frac{\sum_{i \in I} log(RT_i)}{b}$$

and

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i \in I} \left(log(RT_i) - \widehat{\mu}\right)^2}{b-1}}$$

respectively. As we know that sample mean and standard deviation is very sensitive to outliers, the algorithm provides an option to data analysts whether they would like to remove the minimum and maximum RT among the first b correctly answered questions if they believe that there are extreme

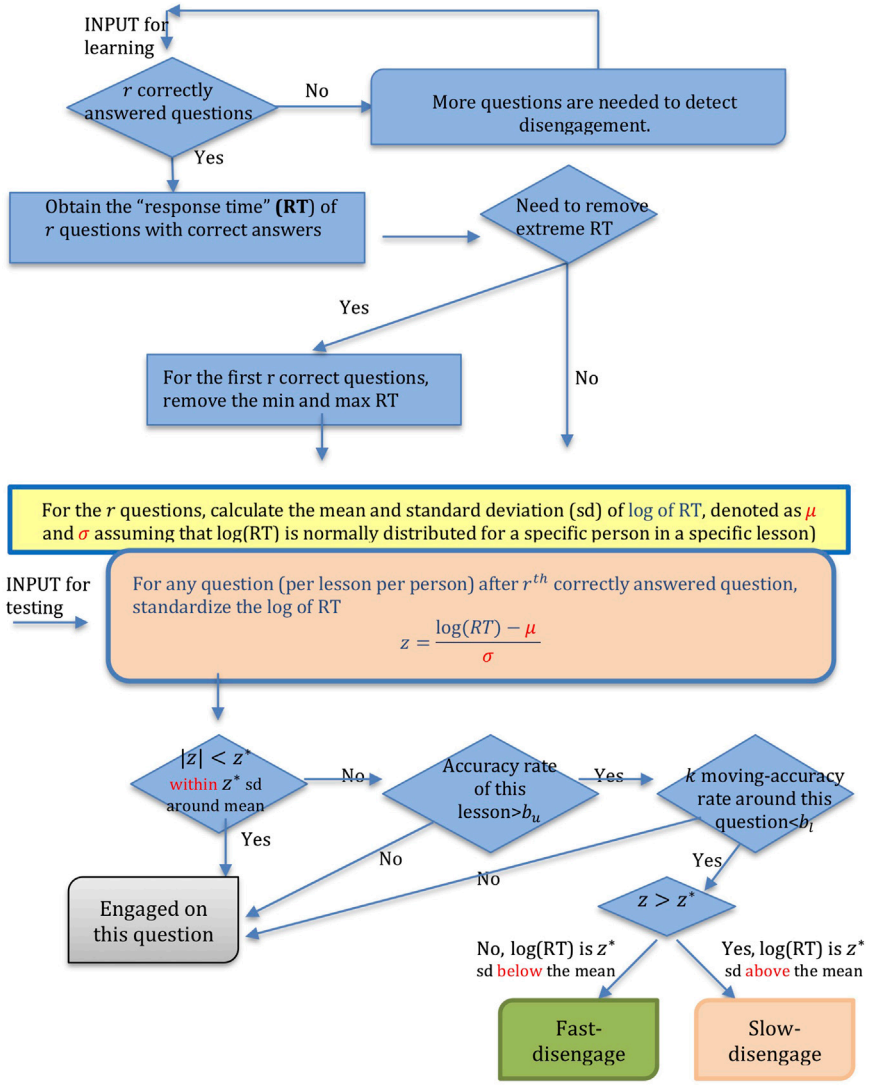

FIGURE 1 | Algorithm Flow Chart of Disengagement Tracing System.

outliers in the $log(RT)$. A student is potentially disengaged at question s if the standardized response time satisfies

$$|z| = \left| \frac{log(RT_s) - \hat{\mu}}{\hat{\sigma}} \right| > z^*,$$

where RT_s is the response time at the s^{th} question and z^* represents the number of standard deviations that the candidate $log(RT)$ departs from the engaged mean of $log(RT)$ to be considered as potential disengagement. A student is slow-disengaged on a question if $z > z^*$, and fast-disengaged if $z < -z^*$. It is known that for normal distribution, 95% should fall within 2 standard deviations and 99.7% should fall within 3 standard deviations. Thus, if we set $z^* = 3$, the probability that a student is falsely tested to be disengaged is only $P(|z| > 3) = 0.03\%$ given the student is actually engaged. Data analysts are free to choose the value of z^* that are appropriate for their study. The choice of z^* should be guided by users' tolerance of false positives. Theoretically, the probability of false positives (i.e. false disengagement) would be 5% if $z^* = 2$. In our analysis, we chose $r = 5$ and $z^* = 3$. Specifically, we computed the mean and standard deviation of the $log(RT)$ of the first five correctly answered questions. It is possible that one question might be answered correctly by accident. To take this into consideration, we dropped the highest (and lowest) reading time before calculating the benchmark statistics. In the current dataset, for the five correctly answered questions, we removed questions with the highest (and lowest) response time and calculated the engaged mean and standard deviation of $log(RT)$ with the remaining three questions. If the student has less than 3 (correctly answered) questions, the system will use the response time of question 2 to 4. In our analysis with AutoTutor-ARC data, a student is suspected to be disengaged on a question (too fast or too slow) if the log of response time on this question is below or above 3 standard deviations from the engaged $log(RT)$.

To have an adaptive DTS, the mean and standard deviation of engaged RT should be personalized for different students in different lessons. This means that 20 s may be an engaged RT for student *X*, but may be too fast to be engaged for student Y on the same question. Learners usually get disengaged for a variety of reasons. They may even get disengaged at different questions in different trials of the same lesson. Furthermore, an individual's reading ability may vary depending on the characteristics of the texts (e.g. difficulty, type) included in each lesson. Because of these sources of variation, the system is required to learn engaged RTs (or reference behaviors) for each learner within each lesson.

Disengagement detection that is only based on response time would lead to a large number of false positives. Some lessons have a small number of "confidence-boosting" questions, which means learners will respond more quickly with high accuracy to these questions than to others. Students may slow down in subsequent questions that are more challenging, which may be falsely detected as disengagement by an DTS that only relies on response time. Other than "abnormal" response time, another important signal of disengagement is that disengaged students usually perform poorly since they are not focusing on the question. If a good student (whose overall performance within the lesson or up to the current question is high, e.g. greater than 80%) responds to a particular question too fast or slow and also answers this question as well as neighbor questions incorrectly in a sequence, there is a high chance that this student is disengaged while working on the particular question. However, if a student performs poorly throughout the entire lesson, DTS should not categorize the questions with abnormal response time and poor performance as disengagement since the student may be struggling with this lesson, but not just disengaged on a few questions. As noted, in this study, DTS only identifies disengaged question-answer observations when assuming that the texts and questions are within the zone of what the student can handle and the student is engaged at the beginning of each lesson. Our targeted disengaged question-answer observations are those with "abnormal" (too fast or too slow) response times, poor local performance, but adequate overall performance. Students with low performance and engagement throughout the entire lesson or study is important also. It is possible that the content of the texts or questions may be too difficult. It is important to note that these questions will not be treated as 'disengaged' by DTS in this study.

A more formal specification of the algorithm may lend clarity. Let X_i be a binary random variable indicating whether the i^{th} question is correctly answered (1: yes, 0: no). Overall performance is the accuracy rate that a student performs in a lesson (or up to current question s), defined as $\frac{\sum_{i=1}^{s} X_i}{s}$. Local performance of a question per participant in a lesson is characterized by moving averages of correctness proportion. The k^{th}-order moving average of s^{th} question is given by $\frac{\sum_{i=s-k}^{s+k} X_i}{2k+1}$. If a student learner's overall performance in the lesson up to s^{th} question is higher than a threshold b_u and k^{th}-order moving average around s^{th} question is below b_l, then this student is detected as 'disengaged'. In this study, we take $k = 1$, $b_u = b_l = 0.5$ and the overall performance is calculated based on all questions in a lesson. The second part of DTS refines the filtering system by not treating well-performed question-answer observations as "disengaged" although students spent abnormal time on these questions. By additionally taking the students' performance into consideration, DTS refines the results from the first part of DTS and largely reduces the false positives in disengagement detection confounded by other factors irrelevant to disengagement. For example, students may spend significantly more response time on a question on new or difficult material. Or a student may struggle the entire lesson and not perform well throughout the lesson (This article aims to detect specific periods where a student gets disengaged, rather than detecting disengaged students. DTS will not treat a student as disengaged when he/she is focused but struggling on this question.).

It is important to reiterate that the proposed DTS algorithm will not handle occurrences when a student is disengaged from the very beginning. These occurrences would not be counted as disengagement (even though they should be) so our predictive algorithm is conservative rather than been generous at detecting disengagement and such observations will dilute the predictive power of the DTS algorithm. It is also important to acknowledge that the algorithm has not yet been validated by self-reports of disengagement, eye tracking, and neurophysiological measures so the precise psychological status of the disengaged observations await further research. That being said, D'Mello and has colleagues (Mills et al., 2017; Faber et al., 2018; D'Mello, 2019) have proposed a decoupling algorithm of disengagement that identifies deviations between a person's self-paced reading times and projected times based on the difficulty of the material, where there is more decoupling when the times are too fast or two slow compared to the projection times; the decoupling algorithms significantly predict self-reported mind-wandering and eye tracking patterns.

Study of Disengaged Question-Answer Observations

The proposed DTS is designed to be a real-time monitoring of disengagement in an intelligent system. Using the predicted engagement/disengagement status for individual questions, we explored the pattern of disengagement in the AutoTutor-ARC as an empirical evaluation of the algorithm (It is important to clarify that DTS was not used during the CSAL AutoTutor study. It was developed after the end of the study.). For each of 252 students, we calculated the proportions of disengaged items, including fast- and slow-disengaged question-answer observations. A k-mean clustering analysis was applied to develop student profiles on proportions of the two types of disengaged question-answer observations. K-means clustering assigns data points into groups by iteratively reassigning and re-averaging the cluster centers until the points have reached convergence (Hartigan and Wong, 1979). Grouping students with similar disengagement patterns could help us to interpret reasons for disengagement within each group of students, and use this information to guide the design of effective interventions to re-engage student users. Fang et al. (2018) performed clustering analysis on the accuracy and response time of the 252 participants and categorized the participants into four groups of adults: higher performers, conscientious readers, under-engaged readers, and

struggling readers. This study compared the clusters of students with different disengagement patterns according to Fang et al.'s four clusters. As a note, Fang et al.'s study removed questions with extreme outliers (i.e. response time was three interquantile range higher than the third quantile).

As an independent evaluation, learning gains were analyzed on subsets of the 252 students who took three standardized tests of comprehension before and after the larger CSAL AutoTutor intervention that included AutoTutor-ARC lessons. Of the 252 participants, 205 took both pre- and post-test of the Woodcock Johnson III Passage Comprehension subtest (Woodcock et al., 2007); 143 took Reading Assessment for Prescriptive Instructional Data (RAPID) Passage Comprehension subtest develop by Lexia Learning (Foorman et al., 2017) and 142 took Reading Inventory and Scholastic Evaluation (RISE) battery developed by ETS (Sabatini et al., 2019). Fang et al. reported that the learning gains in Woodcock Johnson and RAPID tests were highest for conscientious readers, lowest for struggling readers, with higher performing readers and under-engaged readers in between (Fang et al., submitted). It has been shown that readers who invested the time to answer AutoTutor questions with a modicum of accuracy demonstrated significant learning gains on measures of comprehension (Greenberg et al., submitted), which confirms the relationship between intensity of engagement and learning. However, the analyses by Fang et al. (submitted; submitted) were conducted on the aggregate performance and response times of the lessons and items over the 4-month intervention rather focusing on engagement within a particular lesson for a particular student, the focus of the present study.

To investigate whether the learning gain is affected (presumably reduced) by disengagement, we performed paired t-test on the pre- and post-test scores of the three standardized tests after contrasting groups of students with different disengagement patterns at a fine grain level (i.e., students with a high vs. a low proportion of disengaged question-answer observations according to the DTS algorithm). These groups were obtained from the clustering analysis of the 252 participants on proportions of disengaged question-answer observations. It should be noted that these DTS-based clusters are different from the aggregate-based clusters identified by Fang et al. (submitted). We compared the two different types of clusters in this paper. Moreover, we tested the association of learning gain measured by the three standardized tests with the AutoTutor accuracy (i.e. proportions of questions correctly answered by students) of 252 participants. We first separated engaged and disengaged question-answer observations detected by the proposed DTS. For engaged (or disengaged) questions, reading comprehension at post-test was regressed onto the accuracy of engaged (or disengaged) questions adjusted by reading comprehension at pre-test.

RESULTS

Accuracy for Disengaged Versus Engaged Question-Answer Observations in AutoTutor

We applied the proposed DTS algorithm to the data extracted from AutoTutor (67,235 answers to questions from 252 participants in 30 lessons). We identified 16,851 questions with "abnormal" response times, of which 3,082 were disengaged question-answer observations (including 961 fast-disengaged and 2,121 slow-disengaged question-answer observations) among the 252 participants. **Table 2** presents the number of disengaged vs. engaged question-answer observations that were correctly answered. Among 3,082 disengaged question-answer observations, 569 were correctly answered, which represents 18.5% of the total disengaged question-answer observations detected by the proposed DTS algorithm. In contrast, 46,059 (71.8%) of the engaged question-answer observations were answered correctly. To test the association of the correctness and disengagement status in AutoTutor, we ran a generalized linear mixed model by letting the correctness of a question-answer observation as the response variable (1: correct, 0: incorrect) and disengagement status (1: disengaged, 0: engaged) as the predictor, and adding two random terms to adjust the correlated observations due to same student and lesson. It is shown that the odds of answering a question correctly when disengaged is only 8% of the odds when engaged. Quite clearly, when students are disengaged while working on questions in a lesson, their performance on the questions will be significantly lower than the engaged questions (**Table 2**, p-value $< .001$). As discussed earlier, disengagement is one of multiple reasons why students might give wrong answers to a question (e.g., the question is difficult for them, their diligent reasoning is unsuccessful), but we presume that disengagement is a very plausible explanation in a high percentage of the observations. See D'Mello (2019); Millis et al. (2017) in their validation of the decoupling model.

Clusters of Participants and Lessons on Proportions of Disengagement

After aggregating the total number of questions from the lessons, we obtained the frequencies and proportions of disengaged question-answer observations for each of the 252 participants. The total number of questions that a student answered varied from ~ 10 to ~ 500, of which only a very small portion of questions (approximately $3 \sim 9\%$) were disengaged question-answer observations. We identified more questions that were slow-disengaged than fast-disengaged (3.2% vs 1.4%).

Some students tend to have a higher proportion of fast-disengaged question-answer observations, whereas others have more slow-disengaged question-answer observations and yet others are high in both. To address this, a k-mean clustering analysis was performed on groups of students with similar disengagement patterns according to the DTS algorithm. Since students answered a different number of questions, we focused on the proportion (rather than the count) of fast- and slow-disengaged question-answer observations for each participant. The k-mean clustering analysis was implemented in R (version 3.6.0) on the proportions of fast- and slow-disengaged question-answer observations. We clustered the 252 participants into four groups ($k = 4$) according to the "elbow" method by visualizing the plot of "number of clusters" vs. "within groups sum of squares".

TABLE 2 | Number (proportion) of correctness among disengaged vs. engaged question-answer observations.

	Number of questions correctly answered (Correctness rate)	Number of questions incorrectly answered (Incorrectness rate)	Total
Disengaged question-answer observations	569 (18.5%)	2,513 (81.5%)	3,082
Engaged question-answer observations	46,059 (71.8%)	18,094 (28.2%)	64,153

**Linear mixed model: coefficient = −2.56, odds ratio = exp(−2.56) = 0.08, p-value <0.001.*

Figure 2 plots the four clusters of participants with different disengagement patterns. The mean and standard deviation of the proportion of fast- and slow-disengaged observations in each cluster are provided in **Table 3**. The first cluster (red dots in **Figure 2**, labeled *HiFast/HiSlow* for short) represents students with a relatively medium-to-high proportion of fast-disengaged question-answer observations (2%) and a comparatively high proportion of slow-disengaged question-answer observations (7%). The second cluster (deep blue dots in **Figure 2**, labeled *LowFast/HiSlow* for short) includes students with a small proportion of fast-disengaged question-answer observations (1%) and a medium-to-high proportion of slow-disengaged question-answer observations (4%). The third cluster (aqua blue dots in **Figure 2**, labeled *HiFast/LowSlow* for short) represents students with a high proportion of fast-disengaged question-answer observations (3%) and a small proportion of medium-to-high slow-disengaged question-answer observations (2%). The last cluster (green dots in **Figure 2**, labeled *Engaged* for short) represents students with small proportion of fast-disengaged question-answer observations (1%) and small proportion of slow-disengaged question-answer observations (2%). **Figure 2** confirms that the four clusters are visually distinct in the scatterplots. Interestingly, **Figure 2** shows that there are several students with nearly zero fast-disengaged question-answer observations, but a medium-to-high proportion of slow-disengaged observations. It is possible that some of these slow-disengaged observations are not truly disengaged, but rather are instances when the student is encountering difficult questions for them. However, our assumption is that a significant percentage of the questions reflect disengagement because the performance of the students was respectable in the early phase of a lesson.

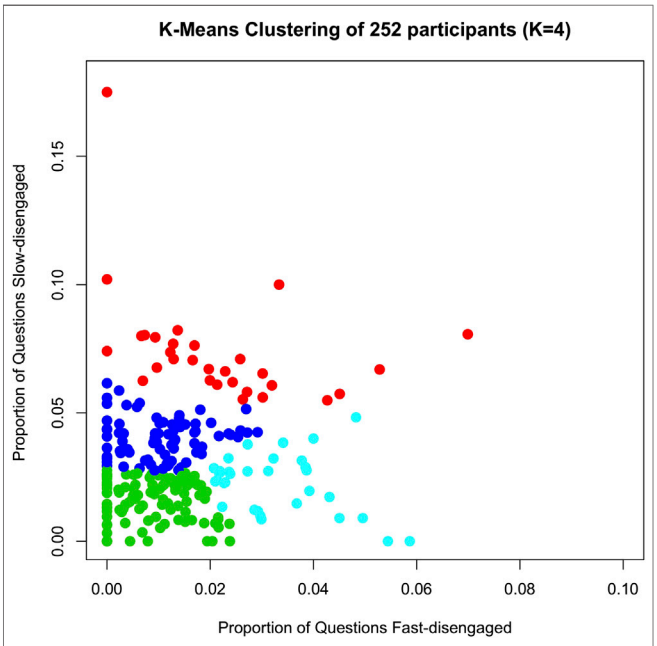

FIGURE 2 | K-mean clustering of 252 participants on the proportion of fast- and slow-disengagement rate. Red dots: students with a medium-to-high proportion of fast-disengaged question-answer observations and high proportion of slow-disengaged question-answer observations (*HiFast/HiSlow*); Deep blue dots: students with a small proportion of fast-disengaged question-answer observations and medium-to-high proportion of slow-disengaged question-answer observations (*LowFast/HiSlow*); Aqua blue dots: students with a high proportion of fast-disengaged question-answer observations and small proportion of slow-disengaged question-answer observations (*HiFast/LowSlow*); Green dots: students with a small proportion of fast-disengaged question-answer observations and small proportion of slow-disengaged question-answer observations (*Engaged*).

The current classification based on local engagement (**Figure 2** and **Table 3**) was compared with the clustering of 252 students in the Fang et al. (2018) study that classified students into four groups based on their accumulated profile over the 4-month intervention. Fang et al. (2018) categorized the 252 participants into four groups: higher performers (fast and accurate), conscientious readers (slow and accurate), under-engaged readers (fast, but lower accuracy) and struggling readers (slow and inaccurate). **Table 4** compares the clusters identified in this study according to the local disengagement patterns with the ones reported in Fang et al. (2018; submitted) that considered the global performance profile. We applied chi-squared test of independence on the overlapped counts of the two sets of clusters (4-by-4 table, **Table 1**) and found a significant association ($\chi^2 = 26.33$, p-value= .002) between the clusters developed by this study and Fang et al. (2018; submitted). According to **Table 4**, a high percentage (52% = 50/97) of "higher performers" are classified as *Engaged* students by DTS, which is higher than "conscientious" and "struggling readers" (42% = 13/31) and much higher than "under-engaged" reader (34% = 32/93). Furthermore, when considering the students with local disengagement (including *HiFast/HiSlow*, *LowFast/HiSlow* and *HiFast/LowSlow*), the conditionalized percentages on on the slow end rather than the fast end were : higher performers (41/47 = 87%), conscientious (7/18 = 39%), struggling (14/18 = 78%), under-engaged (49/61 = 80%); low relative percentages for the

TABLE 3 | Mean and standard deviation (SD) of fast- and slow-disengage proportions for the four clusters of participants in AutoTutor.

Cluster of disengagement from DTS	Mean (SD) of fast-Disengage rate	Mean (SD) of slow-Disengage rate
1 (Red): Disengaged- *HiFast/HiSlow*	0.02 (0.02)	0.07 (0.02)
2 (Deep Blue): Disengaged-*LowFast/HiSlow*	0.01 (0.01)	0.04 (0.01)
3 (Aqua Blue): Disengaged- *HiFast/LowSlow*	0.03 (0.01)	0.02 (0.01)
4 (Green): *Engaged*	0.01 (0.01)	0.02 (0.01)

HiFast/HiSlow—disengaged students with a medium-to-high proportion of fast-disengaged question-answer observations and high proportion of slow-disengaged question-answer observations;
LowFast/HiSlow—disengaged students with a small proportion of fast-disengaged question-answer observations and medium-to-high proportion of slow-disengaged question-answer observations;
HiFast/LowSlow—disengaged students with a high proportion of fast-disengaged question-answer observations and small proportion of slow-disengaged question-answer observations;
Engaged—students with a small proportion of fast-disengaged question-answer observations and small proportion of slow-disengaged question-answer observations.

conscientious readers is unexpected, but perhaps can be attributed to the relatively small number of observations.

The major discrepancy between the two clustering approaches can be attributed to the fact that DTS was developed to detect disengaged question-answer observations, rather than disengaged students. Thus, DTS only checks the accuracy of answers to a question locally (i.e. accuracy of neighbored questions), not globally (e.g. accuracy within lessons that accumulated over the 4-month intervention). In our study, a student is considered to be disengaged while working on a question if his/her performance on this (and neighbored) questions is lower than their global performance. If a student has a low accuracy throughout the entire lesson, DTS will count these question-answer observations as *Engaged*. In contrast, Fang et al. categorized readers with low global accuracy to "under-engaged."

The next analysis computed the proportion of fast- and slow-disengaged question-answer observations among the 252 participants within each of the 30 lessons separately. **Figure 3** shows these results for the 30 lessons in the approximate order that the lessons occurred in the curriculum (there were small deviations in the sequence over the course of the intervention). **Figure 3** shows that the proportions of fast- and slow-disengaged observations differed among the 30 lessons. Some lessons have a larger proportion of slow-disengaged question-answer observations than others. For example, lesson #04-Word Parts and #07-Learning New Words clearly have a higher proportion of fast-disengaged question-answer observations compare to lesson #13-A Personal Story and #14-Connecting Ideas. To better understand which lessons are more (or less) likely to get students disengaged, with the fast- and slow-disengagement proportions in each lesson, we clustered the 30 lessons in terms of their disengagement pattern using k-mean clustering analysis. Exploring the disengagement pattern across lessons would provide AutoTutor designers critical information and guidance to adjust the difficulty levels of content and/or enhance the display interfaces of questions in lessons to diminish or prevent disengagement. These results are presented in Appendix A. Three groups of lessons were chosen. This first group contains lessons, such as "Text

TABLE 4 | Comparisons of clusters of 252 participants.

Clusters according to local disengagement pattern Identified by DTS	Clusters reported in Fang et al. (2018; submitted) over 30 lessons			
	Higher performers	Conscientious readers	Struggling readers	Under-engaged reader
	Count (%)	Count (%)	Count (%)	Count (%)
1 (Red): Disengaged-*HiFast/HiSlow*	8 (8%)	3 (10%)	3 (10%)	16 (17%)
2 (Deep Blue): Disengaged-*LowFast/HiSlow*	33 (34%)	4 (13%)	11 (35%)	33 (35%)
3 (Aqua Blue): Disengaged-*HiFast/LowSlow*	6 (18%)	11 (35%)	4 (13%)	12 (13%)
4 (Green): *Engaged*	50 (52%)	13 (42%)	13 (42%)	32 (34%)
Total (%)	97 (100%)	31 (100%)	31 (100%)	93 (100%)
Chi-squared test of independence: χ^2 = 26.33, *p*-value= 0.002				

HiFast/HiSlow—disengaged students with a medium-to-high proportion of fast-disengaged question-answer observations and high proportion of slow-disengaged question-answer observations;
LowFast/HiSlow—disengaged students with a small proportion of fast-disengaged question-answer observations and medium-to-high proportion of slow-disengaged question-answer observations;
HiFast/LowSlow—disengaged students with a high proportion of fast-disengaged question-answer observations and small proportion of slow-disengaged question-answer observations;
Engaged—students with a small proportion of fast-disengaged question-answer observations and small proportion of slow-disengaged question-answer observations.

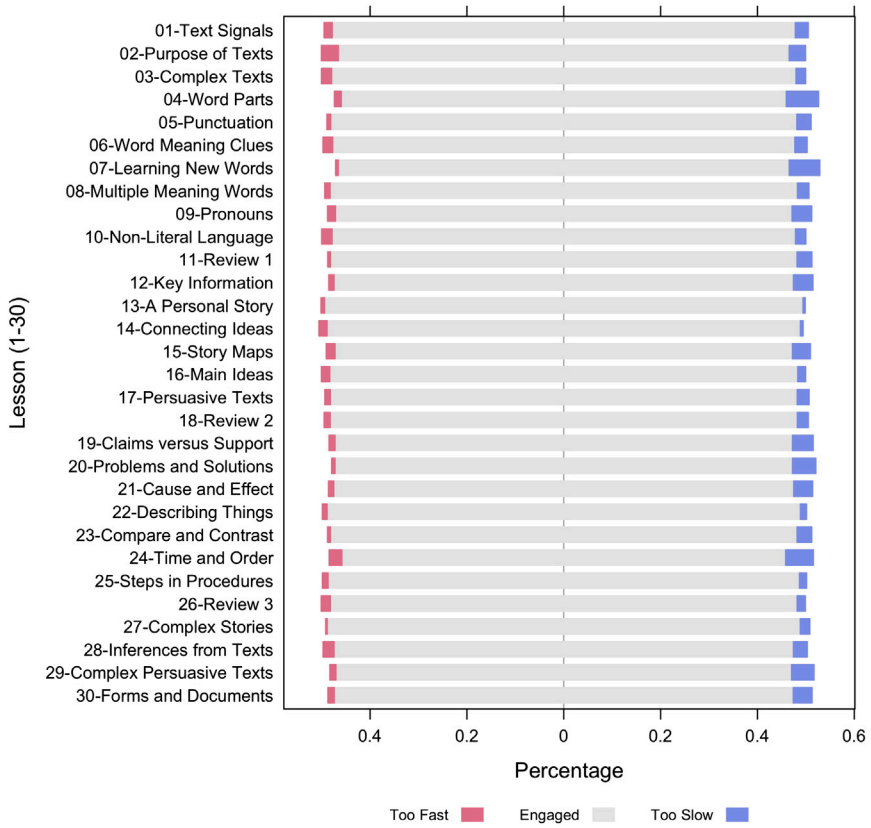

FIGURE 3 | Proportions of disengaged question-answer observations from 252 participants in 30 lessons.

Signals", "Purpose of Texts", "Inferences from Texts", has balanced proportions of fast- and slow-disengaged question-answer observations. The second group of lessons have higher proportions of slow-disengaged question-answer observations. Lessons in the second cluster include "Claims vs. Support", "Cause and Effect", "Persuasive Texts", which are more advanced and difficult topics and lead to an increased slow-disengage. The proportion of both fast- and slow-disengage is low in the third group of lessons.

Proportion of Disengaged Question-Answer Observations for Different Difficulty Levels and Theoretical Levels

A subset of the lessons have one or two texts with conversation-based questions woven into the lessons. Eleven of the lessons have multi-sentence texts that branched during the course of the lessons. For each of these lessons with branching texts, the AutoTutor system starts with a medium difficulty text with 8–12 questions and then branches to an easy or hard text, depending on the student's performance on the questions in the medium difficulty texts. A second set of nine lessons provide one medium level text with 10–20 questions woven into the conversation about the text. A third set of 10 lessons focused on single words or sentences rather than multi-sentence texts. These lessons had 10–30 questions that were scaled on easy, medium or difficult levels. When considering all 30 lessons, the questions at the medium difficulty level constituted the majority of questions. Since some lessons contain questions of different difficulty levels, we evaluated the proportion of fast- and slow-disengaged items stratified by difficulty levels of questions for 252 participants in the 30 lessons. **Figure 4** provides the bar chart with the percentage of disengaged question-answer observations at different difficulty levels. Easy questions had a slightly larger proportion of fast-disengage compared to the other two types of questions. This can be explained by the plausible possibility that some students are bored by the easy questions and quickly click the answers. **Figure 4** also indicates that the proportion of slow-disengaged observations is the highest in hard questions, which is very reasonable since students may need more time to work on hard questions; students may give up on the hard questions and get disengaged. In order to statistically assess whether the differences are reliable, we conducted a generalized linear mixed model by setting the disengagement status (1: disengaged, 0: engaged) as the response variable, level of difficulty (easy/medium/hard) as the predictor variable and adding two random terms to adjust for variability among students and lessons. The results confirmed that students tend to be disengaged more often on hard in comparison to easy questions (odds ratio = 1.5, $p < .001$).

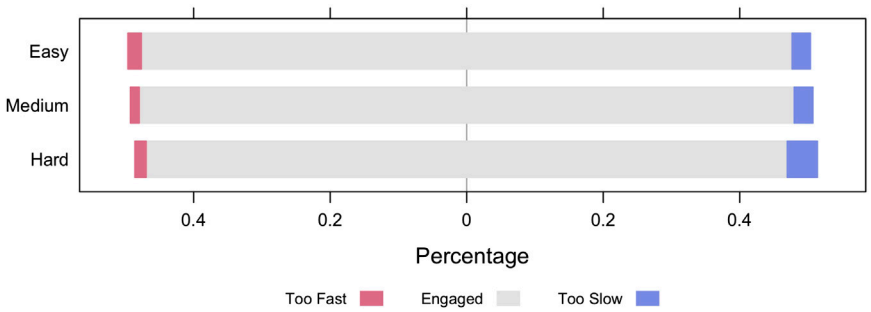

FIGURE 4 | Disengagement rate (Too Fast: fast-disengaged; Too Slow: slow-disengaged) for questions of different difficulty levels.

The AutoTutor lessons were also scaled on four theoretical levels: Words (W), textbase (TB), Referential Situation Model (RSM), and Rhetorical Structure (RS), based on Graesser and McNamara's (2011) multilevel theoretical framework. A description of these theoretical levels is provided in Study of *Disengaged Question-Answer Observations*. For each person and lesson, we calculated the proportion of disengaged question-answer observations. To test whether the disengagement rate of lessons from one theoretical level is different from another, a linear mixed model was conducted while adjusting the correlated observations due to the same student. The results revealed that lessons in the Word (W) theoretical level had the highest disengagement rate (1% higher than RS with $p = .005$, 1.1% higher than TB with $p = .014$ and 1.7% higher than RSM with $p < .001$). However, the differences were surprisingly small and not different for the fast-vs. slow-disengaged items.

Association With Learning Gains From Three Standardized Tests of Comprehension

Comprehension was evaluated by three standardized tests (Woodcock Johnson III Passage Comprehension, RISE and RAPID). There was a pretest before the 4-month intervention and a posttest at the end of it. Learning gain is calculated by the difference between pre- and post-test. To assess whether disengagement has an effect on learning gains in AutoTutor, we separated disengaged question-answer observations from the engaged ones and tested the association between learning gains from three standardized tests of comprehension and the accuracy rate (aggregated from all lessons) in AutoTutor on disengaged and engaged observations respectively. Regression analyses were conducted on the learning gains in the comprehension tests as a function of the AutoTutor intervention with engaged vs. disengaged question-answer observations. These results are presented in **Table 5**. Learning gains on the three standardized tests were significantly predicted by the accuracy in AutoTutor on engaged question-answer observations, but were not significant on disengaged question-answer observations. For example, when accuracy rate of engaged questions increases by one unit, the mean learning gains on Woodcock Johnson III Passage Comprehension test increase by 0.56 (p-value < 0.001). However, the change in accuracy rate of disengaged questions is not statistically significantly associated with learning gains on Woodcock Johnson III Passage Comprehension test.

DISCUSSION AND SUMMARY

This paper provides a disengagement tracking system (DTS) with an intelligent algorithm to monitor students' disengagement based on their response time and performance on each question during their learning process in AutoTutor. A variety of approaches have been applied to predict and track disengagement in intelligent tutoring systems (Allen et al., 2016; Bixler and D'Mello, 2013; D'Mello and Graesser, 2012). Existing disengagement/engagement detection methods mainly predict disengagement/engagement by applying supervised learning approaches using self-reported mind-wandering (Bosch and Dmello, 2019; Mills and D'Mello, 2015; Millis et al., 2017). These methods are not suitable for personalized and concurrent disengagement detection. Tracking students' disengagement promptly would allow personalized interactions at appropriate times in order to re-engage students.

The proposed DTS consists of two steps. In the first step, the algorithm learns a student's baseline response time from his/her first 3 ~ 5 well-performed questions in a specific lesson and creates a personalized reference of response time. This first step rests on the plausible premise that the student is engaged at the beginning of a lesson. A student is suspected to be "disengaged" on a question if the response time on a question abnormally deviates from the baseline, which is expected to be more prevalent after the initial phase of a lesson. In the second step, the algorithm checks all the 16,851 candidate disengaged question-answer observations and marks those with good overall performance in a lesson (proportion of correctly answered questions is higher than a threshold) but poor local accuracy (proportion of correctness rate in the neighbor questions but not the target question is lower than a threshold) as disengaged question-answer observations. The proposed method is derived from the time and accuracy of data in log files and does not require any self-reported reports from the participants or physiological measures of engagement. Moreover, the DTS algorithm can detect disengagement within small time spans

TABLE 5 | Predicted learning gains from pretest to posttest on three standardized tests (RISE, RAPID and Woodcock-Johnson passage comprehension) from engaged vs. disengaged question-answer observations.

Types of pre- and post-tests	Engaged question-answer observations predicting learning gains (*p*-value)	Disengaged question-answer observations predicting learning gains (*p*-value)
Woodcock Johnson	0.56 (0.007)[a]	−0.06 (0.662)
RISE	2.26 (< 0.001)[b]	0.41 (0.130)
RAPID	0.58 (< 0.001)[b]	−0.01 (0.897)

[a]*indicates that the p-value < 0.01.*
[b]*indicates that the p-value < 0.001.*

of a minute or two rather than after a lesson or dozens of lessons have been completed. For instance, if a student is disengaged starting from the ninth question, the earliest time that the algorithm would be able to capture it is after the student completed the 11th question. The proposed algorithm offers low computational burden and can be included *in vivo* as a performance monitoring algorithm within an intelligent tutoring system.

Our study of disengaged question-answer observations in AutoTutor that were identified by DTS is consistent with the claim that disengaged observations have substantially lower accuracy on AutoTutor items whereas engaged observations high performance. This is a confirmation of the internal validity of the algorithm. Evidence of external validity was also confirmed in analyses of learning gains on comprehension skills that were measured by independent psychometric tests (Woodcock et al., 2007; Foorman et al., 2017; Sabatini et al., 2019). Learning gains on these tests were predicted by the accuracy rate of engaged question-answer observations in AutoTutor but not the disengaged observations. These two lines of evidence suggest that the evaluation and tuning of AutoTutor or other ITSs could benefit from analyzing the engagement profiles reflected in question-answer observations and that the DTS is a promising algorithm to detect disengagement.

Disengagement detection and monitoring is of course important for improving learning in conventional learning contexts as well as intelligent tutoring systems (Csikszentmihalyi, 1990; D'Mello and Graesser, 2012; Larson and Richards, 1991; Mann and Robinson, 2009; Millis et al., 2017; Pekrun et al., 2010; Pekrun and Stephens, 2012). A few ITS studies have been conducted with personalized interventions to prevent or interrupt disengaging behaviors and guide an individual learner back on track (Bosch et al., 2016; D'Mello, 2019; D'Mello and Graesser, 2012, 2012; Lane, 2015; Monkaresi et al., 2017; Woolf et al., 2010). Feedback from the proposed disengagement tracking system can elucidate factors that lead to distractions or impetuous responding. Was it the question or content difficulty or low interest in the material, poor pacing, lack of razzle dazzle, or perceived value of the learning experience? ITS can also be designed to engage the off-track student at the right time. For example, once the disengagement is identified, a conversational agent or pop-up window can express one or more of the following messages: It seems like you may be distracted. Do you need a break? Would you like to continue to learn more about XX? Alternatively, the ITS could present more difficult or easy material to optimize students' zone of attention and learning (Graesser et al., 2016a). These interventions will hopefully encourage students to turn their attention back to the lesson. The false-positives and false-negatives generated by this DTS may or may not be problematic, depending on how DTS integrates with adaptive elements of the ITS. While this is beyond the scope of this paper, the optimal system response to disengagement may, for example, align with the optimal system response to slow engagement on difficult items. To the extent that optimal system responses overlap, DTS errors are not problematic. In cases where the appropriate system response should differ, these offer opportunities to improve DTS.

There are a number of limitations in this study that call for follow-up research. First, we assumed that the log-transformed response time follows a normal distribution, and hence an "abnormal" response time can be identified if a log-transformed response time falls outside of z^* standard deviation of its mean. The resulting distributions of the log-transformed response times confirmed that the distributions were normal. However, some data sets might not exhibit a normal distribution. To accommodate any severely skewed or heavy-tailed distributions, the proposed method can be revised by replacing the mean and standard deviation with more robust alternatives, e.g. median and median absolute deviation (MAD) as suggested by (Miller, 1991; Leys et al., 2013). Thus, a student will be suspected to be disengaged on a question if the response time on this question is below or above three MAD from the median response time of engaged items. These possibilities can be explored in future research.

Second, the DTS algorithm assumes that questions in a lesson are similar/interchangeable in terms of the lesson content and difficulty. **Figure 3** and Appendix A display the variations among the 30 lessons. Somewhat surprisingly, there were very small and primarily nonsignificant differences when comparing the theoretical levels of the lessons (words, textbase, situation model, rhetorical structure). Our study revealed that the proportion of slow-disengaged observations is higher in the comparatively hard questions (see **Figure 4**). As discussed earlier, the literature has confirmed that disengagement and mindwandering increase with the difficulty of expository reading materials (D'Mello, 2019; Feng et al., 2013; Miller, 1991; Mills et al., 2017). In our future studies, we may improve the DTS by adding a factor that annotates text/item difficulty or difficulty transitions to prevent falsely discovering slow-disengagement when the materials given to a student branches to harder materials.

Third, the algorithm does not detect situations when the student is disengaged from the material at the beginning of the lesson. For the DTS to be meaningfully applied to AutoTutor, we assume that texts/questions given to students are suitable for them and have some modicum of value and/or interest. This is a plausible initial assumption because the lessons focus on subject matters that have value for struggling adult

readers (e.g., comprehending a rental agreement or a job form) or are interesting to adults. Hence, students presumably start out engaged in most of the questions and may be disengaged on a number of questions some time later. If a student is disengaged in the beginning of a lesson or disengaged from most of the questions, DTS would need to be adjusted with a different algorithm to improve its predictions.

Fourth, there are a number of other situations that the DTS algorithm would need to be modified to handle. The proposed algorithm does not consider any intervention to re-engage the students. DTS needs to be adjusted if any intervention action is taken after a disengaged question is detected. If users encounter frequent technical issues in the early/testing stage of a new ITS system, the data should take that into consideration. DTS run the risk if identifying "false alarms in disengagement" or "misses in disengagement observations" if the questions at the early phase of a lesson are unusual and fail to calibrate their performance when engaged.

In summary, DTS provides an algorithm that can automatically predict/monitor disengaged behaviors in other learning environments that collect self-paced responses to question-answer items during training. It was designed for, but not limited to, the AutoTutor-ARC system. It can be tailored to fit any ITSs. In the proposed algorithm, only the response time and accuracy of each question are utilized to predict disengagement since they are the only relevant items that are recorded by AutoTutor. If other predictors or measurements, such as item difficulty, self-reported engagement or student's gaze patterns captured by a commercial eye tracker are available in different intelligent tutoring systems, they can be incorporated into the proposed DTS with appropriate modifications to the proposed algorithm. These other sources of data can also be used to validate the DTS algorithm. Of course, these other measures may be difficult or impossible to collect when scaling up a learning system in the real world.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: LearnSphere, https://datashop.memphis.edu/ProjectPermissions?id=76.

AUTHOR CONTRIBUTIONS

SC, AG, YF, and GS contributed to the development of research method and writing of the first draft of the manuscript. AG, JS, DG, and JF contributed to the design of AutoTutor. SC, YF, and GS contributed to data preprocessing and conducted statistical analysis. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

This study is supported by the Institute of Education Sciences, US Department of Education, through Grants R305C120001 and R305A200413, the National Science Foundation Data Infrastructure Building Blocks program under Grant No. ACI-1443068 and the National Science Foundation under the award The Learner Data Institute with Grant No. 1934745.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.595627/full#supplementary-material.

REFERENCES

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., and McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: the essay, the writer, and keystrokes. In *Proceedings of the sixth international conference on learning analytics & knowledge*, (New York, NY, USA: Association for Computing Machinery), LAK '16, 114–123. doi:10.1145/2883851.2883939

Baker, R., Corbett, A., Roll, I., and Koedinger, K. (2008). Developing a generalizable detector of when students game the system. *User Model. User-Adapted Interact* 18, 287–314. doi:10.1007/s11257-007-9045-6

Beck, J. E. (2005). Engagement tracing: using response times to model student disengagement. In *Proceedings of the 2005 conference on artificial intelligence in education: supporting learning through intelligent and socially informed technology*, Editors B. B. Chee-Kit Looi, Gord. McCalla, and J. Breuker (Amsterdam, The Netherlands: IOS Press), 88–95.

Bixler, R., and D'Mello, S. (2013). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. in *Proceedings of the 2013 international conference on intelligent user interfaces* (New York, NY, USA: Association for Computing Machinery), *IUI '13*, 225–234. doi:10.1145/2449396.2449426

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., and Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Trans. Interact. Intell. Syst.* 6. doi:10.1145/2946837

Bosch, N., and Dmello, S. (2019). Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing* 1. doi:10.1109/TAFFC.2019.2908837

Bulathwela, S., Pérez-Ortiz, M., Lipani, A., Yilmaz, E., and Shawe-Taylor, J. (2020). Predicting engagement in video lectures. Available at:https://arxiv.org/abs/2006.00592

Cain, K. (2010). *Reading development and difficulties*, Vol. 8 Hoboken, NJ: John Wiley & Sons.

Chen, M., and Kizilcec, R. F. (2020). "Return of the student: predicting re-engagement in mobile learning." in *Proceedings of the thirteenth international Conference on educational data mining. July 10-13, 2020.*

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience* (New York: Harper Collins).

D'Mello, S., and Graesser, A. (2012). Autotutor and affective autotutor: learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems* 2, 1–39.

D'Mello, S. K., and Graesser, A. (2012). Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies* 5, 304–317. doi:10.1109/TLT.2012.10

D'Mello, S., Olney, A., Williams, C., and Hays, P. (2012). Gaze tutor: a gaze-reactive intelligent tutoring system. *Int. J. Hum. Comput. Stud.* 70, 377–398. doi:10.1016/j.ijhcs.2012.01.004

D'Mello, S. (2019). What we think about when we learn. in *Deep comprehension: multidisciplinary approaches to understanding, enhancing, and measuring comprehension*, Editors J. M. K. Millis, D.L. Long, and K. Wiemer (New York, NY: Routledge).

Danielle, S. M., Tenaha, P. O., Rachel, M. B., and Yasuhiro, O. (2006). Improving adolescent students' reading comprehension with istart. *J. Educ. Comput. Res.* 34, 147–171. doi:10.2190/1RU5-HDTJ-A5C8-JVWE

Faber, M., Bixler, R., and D'Mello, S. (2018). An automated behavioral measure of mind wandering during computerized reading. *Behav. Res. Methods* 50, 134–150. doi:10.3758/s13428-017-0857-y

Fang, Y., Shubeck, K., Lippert, A., Chen, Q., Shi, G., Feng, S., et al. (2018). Clustering the learning patterns of adults with low literacy interacting with an intelligent tutoring system. in *Proceedings of the 11th international conference on educational data mining*, Editors. K. Boyer and M. Yudelson (Buffalo, NY: Educational Data Mining Society), 348–354.

Feng, S., D'Mello, S. K., and Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review* 20, 586–592. doi:10.3758/s13423-012-0367-y

Foorman, B. R., Petscher, Y., and Schatschneider, C. (2017). Technical manual for Lexia RAPID assessment version 3.0: grades 3-12 *(Lexia learning)*. Available at: http://www.lexialearningresources.com/RAPID/RAPID_TechnicalK2.pdf

Graesser, A. C., Baer, W., Feng, S., Walker, B., Clewley, D., Hays, D. P., et al. (2016a). Emotions in adaptive computer technologies for adults improving reading. in *Emotions, technology, design, and learning* (Amsterdam, Netherlands: Elsevier). 3–25.

Graesser, A. C., Cai, Z., Baer, W. O., Olney, A., Hu, X., Reed, M., et al. (2016b). Reading comprehension lessons in autotutor for the center for the study of adult literacy. in *Adaptive educational technologies for literacy instruction*, Editors S. Crossley and D. McNamara (New York, NY: Taylor & Francis Routledge), 288–293.

Graesser, A. C., Cai, Z., Morgan, B., and Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Comput. Hum. Behav.* 76, 607–616. doi:10.1016/j.chb.2017.03.041

Graesser, A. C. (2016). Conversations with autotutor help students learn. *Int. J. Artif. Intell. Educ.* 26, 124–132. doi:10.1007/s40593-015-0086-4

Graesser, A. C., Li, H., and Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Curr. Dir. Psychol. Sci.* 23, 374–380. doi:10.1177/0963721414540680

Graesser, A. C., and McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science* 3, 371–398. doi:10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., Conley, M. W., and Olney, A. (2012). Intelligent tutoring systems. in *APA educational psychology handbook, Vol 3: Application to learning and teaching*. Editors K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, and H. L. Swanson (Worcester, MA: American Psychological Association), 451–473.

Hartigan, J. A., and Wong, M. A. (1979). Algorithm as 136: a k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 100–108. doi:10.2307/2346830

Hockey, R. (2013). *The psychology of fatigue: work, effort and control* (Cambridge, UK: Cambridge University Press).

Holden, R., and Passey, J. (2010). Socially desirable responding in personality assessment: not necessarily faking and not necessarily substance. *Pers. Indiv. Differ.* 49, 446–450. doi:10.1016/j.paid.2010.04.015

Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J. R., et al. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Model. User-Adapted Interact* 29, 821–867. doi:10.1007/s11257-019-09228-5

Jackson, G. T., and McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *J. Educ. Psychol.* 105, 1036–1049. doi:10.1037/a0032580

Johnson, W. L., and Lester, J. C. (2016). Face-to-face interaction with pedagogical agents, twenty years later. *Int. J. Artif. Intell. Educ.* 26, 25–36. doi:10.1007/s40593-015-0065-9

Lane, H. C. (2015). Enhancing informal learning experiences with affect-aware. in *The oxford handbook of affective computing*, Editors. J. G. R. A. Calvo, S. K. D'Mello, and A. Kappas (New York, NY: Oxford University Press).

Larson, R. W., and Richards, M. H. (1991). Boredom in the middle school years: blaming schools versus blaming students. *Am. J. Educ.* 99, 418–443. doi:10.1086/443992

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* 49, 764–766. doi:10.1016/j.jesp.2013.03.013

Lippert, A., Shubeck, K., Morgan, B., Hampton, A., and Graesser, A. (2020). Multiple agent designs in conversational intelligent tutoring systems. *Technol. Knowl. Learn.* 25, 443–463. doi:10.1007/s10758-019-09431-8

Mann, S., and Robinson, A. (2009). Boredom in the lecture theatre: an investigation into the contributors, moderators and outcomes of boredom amongst university students. *Br. Educ. Res. J.* 35, 243–258. doi:10.1080/01411920802042911

Miller, J. (1991). Reaction time analysis with outlier exclusion: bias varies with sample size. *Q. J. Exp. Psychol.* 43, 907–912. doi:10.1080/14640749108400962

Millis, K., Forsyth, C., Wallace, P., Graesser, A. C., and Timmins, G. (2017). The impact of game-like features on learning from an intelligent tutoring system. *Technol. Knowl. Learn.* 22, 1–22. doi:10.1007/s10758-016-9289-5

Millis, K. K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., and Halpern, D. F. (2011). "Operation aries!: a serious game for teaching scientific inquiry."in *Serious Games and edutainment applications*. Editors M. Ma, A. Oikonomou, and L. Jain (London, UK: Springer).

Mills, C., and D'Mello, S. (2015). "Toward a real-time (day) dreamcatcher: sensor-free detection of mind wandering during online reading." in *Proceedings of the 8th international conference on educational data mining*. Editors O. Santos, J. Boticario, and C. Romero, and Others (Madrid, Spain: International Educational Data Mining Society), 69–76.

Mills, C., Graesser, A., Risko, E., and D'Mello, S. (2017). Cognitive coupling during reading. *J. Exp. Psychol. Gen.* 146, 872.

Monkaresi, H., Bosch, N., Calvo, R. A., and D'Mello, S. K. (2017). Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing* 8, 15–28. doi:10.1109/TAFFC.2016.2515084

Nye, B. D., Graesser, A. C., and Hu, X. (2014). Autotutor and family: a review of 17 years of natural language tutoring. *Int. J. Artif. Intell. Educ.* 24, 427–469. doi:10.1007/s40593-014-0029-5

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., and Perry, R. P. (2010). Boredom in achievement settings: exploring control–value antecedents and performance outcomes of a neglected emotion. *J. Educ. Psychol.* 102, 531. doi:10.1037/a0019243

Pekrun, R., and Stephens, E. J. (2012). Academic emotions. In *APA educational psychology handbook, Vol. 2: Individual differences and cultural and contextual factors*. Editors K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, and M. Zeidner (Worcester, MA: American Psychological Association), 3–31.

Perfetti, C. (2007). Reading ability: lexical quality to comprehension. *Sci. Stud. Read.* 11, 357–383. doi:10.1080/10888430701530730

Rayner, K., Foorman, B., Perfetti, C., Pesetsky, D., and Seidenberg, M. (2001). How psychological science informs the teaching of reading. *Psychol. Sci. Publ. Interest* 2, 31–74. doi:10.1111/1529-1006.00004

Sabatini, J., Weeks, J., O' Reilly, T., Bruce, K., Steinberg, J., and Chao, S.-F. (2019). *SARA Reading Components Tests, RISE forms: technical adequacy and test design*, 3rd edition (Research Report No. RR-19-36) (Princeton, NJ: Educational Testing Service).

Shi, G., Lippert, A. M., Shubeck, K., Fang, Y., Chen, S., Pavlik, P., et al. (2018). Exploring an intelligent tutoring system as a conversation-based assessment tool for reading comprehension. *Behaviormetrika* 45, 615–633. doi:10.1007/s41237-018-0065-9

Smallwood, J., and Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annu. Rev. Psychol.* 66, 487–518. doi:10.1146/annurev-psych-010814-015331

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., et al. (2010). "Ada and grace: toward realistic and engaging virtual museum guides." in *Intelligent Virtual Agents. IVA 2010. Lecture Notes in Computer Science*. Editors J. llbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova (Berlin, Germany: Springer), 286–300. doi:10.1007/978-3-642-15892-6_30

Woodcock, R. W., McGrew, K. S., and Mather, N. (2007). *Woodcock-johnson III normative update complete*. (Rolling Meadows, IL: Riverside Publishing).

Woolf, B., Arroyo, I., Muldner, K., Burleson, W., Cooper, D., Dolan, R., et al. (2010). "The effect of motivational learning companions on low achieving students and students with disabilities." in *Intelligent tutoring systems*. Editors V. Aleven, J. Kay, and J. Mostow (Berlin, Germany: Springer), 327–337. doi:10.1007/978-3-642-13388-6_37

Woolf, B. P. (2010). *Building intelligent interactive tutors: student-centered strategies for revolutionizing e-learning*. (Burlington, MA: Morgan Kaufmann).

Zapata-Rivera, D., Jackson, G., and Katz, I. (2015). Authoring conversation-based assessment scenarios. in *Design recommendations for intelligent tutoring systems*. (Orlando, FL: U.S. Army Research Laboratory), Vol. 3, 169–178.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

SYSTEMATIC REVIEW
published: 09 December 2021
doi: 10.3389/frai.2021.748454

Persuasive Apps for Sustainable Waste Management: A Comparative Systematic Evaluation of Behavior Change Strategies and State-of-the-Art

Makuochi Nkwo[1]*, *Banuchitra Suruliraj*[2] and *Rita Orji*[2]*

[1]Department of Computer Science, Ebonyi State University, Abakaliki, Nigeria, [2]Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

OPEN ACCESS

Edited by:
Amon Rapp,
University of Turin, Italy

Reviewed by:
Arghir-Nicolae Moldovan Moldovan,
National College of Ireland, Ireland
Chien-Sing Lee,
Sunway University, Malaysia

***Correspondence:**
Makuochi Nkwo
makuonkwo@gmail.com
Rita Orji
Rita.orji@dal.ca

Specialty section:
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

Received: *28 July 2021*
Accepted: *21 October 2021*
Published: *09 December 2021*

Citation:
Nkwo M, Suruliraj B and Orji R (2021) Persuasive Apps for Sustainable Waste Management: A Comparative Systematic Evaluation of Behavior Change Strategies and State-of-the-Art. Front. Artif. Intell. 4:748454. doi: 10.3389/frai.2021.748454

With the proliferation of ubiquitous computing and mobile technologies, mobile apps are tailored to support users to perform target behaviors in various domains, including a sustainable future. This article provides a systematic evaluation of mobile apps for sustainable waste management to deconstruct and compare the persuasive strategies employed and their implementations. Specifically, it targeted apps that support various sustainable waste management activities such as personal tracking, recycling, conference management, data collection, food waste management, do-it-yourself (DIY) projects, games, etc. The authors who are persuasive technology researchers retrieved a total of 244 apps from App Store and Google Play, out of which 148 apps were evaluated. Two researchers independently analyzed and coded the apps and a third researcher was involved to resolve any disagreement. They coded the apps based on the persuasive strategies of the persuasive system design framework. Overall, the findings uncover that out of the 148 sustainable waste management apps evaluated, primary task support was the most employed category by 89% (n = 131) apps, followed by system credibility support implemented by 76% (n = 112) apps. The dialogue support was implemented by 71% (n = 105) apps and social support was the least utilized strategy by 34% (n = 51) apps. Specifically, *Reduction* (n = 97), *personalization* (n = 90), *real-world feel* (n = 83), *surface credibility* (n = 83), *reminder* (n = 73), and *self-monitoring* (n = 50) were the most commonly employed persuasive strategies. The findings established that there is a significant association between the number of persuasive strategies employed and the apps' effectiveness as indicated by user ratings of the apps. How the apps are implemented differs depending on the kind of sustainable waste management activities it was developed for. Based on the findings, this paper offers design implications for personalizing sustainable waste management apps to improve their persuasiveness and effectiveness.

Keywords: sustainability, waste management, mobile apps, persuasive strategies, behavior change, systematic review

INTRODUCTION

Persuasive technology is a sub-discipline of Human–Computer Interaction (HCI) that has evolved over the last 15 years. However, in recent years, the personalization of persuasive technologies has generated growing interest in the application of persuasion to technology design. Advances in smart and mobile technologies have created opportunities and shaped the way that billions of users worldwide connect and socialize with one another (Gu, 2019), learn new ways of doing things (Orji, 2017), and perform target behaviors (Istudor and Gheorghe Filip, 2014). As a result, mobile solutions such as apps and games have become attractive channels to deliver personalized and socially responsible interventions. Many of these apps and games are environmentally related and help to encourage positive individual and communal actions toward the realization of the United Nations (UN) Sustainable Development Goals (SDGs) as it concerns environmental protection and sustainability programs (such as global climate change action plans, etc.), as well as promote the health and wellbeing of the people (Nkwo et al., 2020). Specifically, these mobile sustainability apps are effective in encouraging energy conservation (Gustafsson, 2010), water preservation (Paay et al., 2013), waste management (Nkwo et al., 2018), and so on.

Sustainable waste management plays a significant role in ensuring the health and wellbeing of the people. Efforts by governments and stakeholders around the world, aimed at ensuring that citizens adopt appropriate waste disposal behaviors, have been largely ineffective (Thieme et al., 2012; Nkwo et al., 2018), hence the calls for new approaches, which can be achieved *via* the combined powers of technologies and persuasive strategies. As a result, there are increasing interests and investments in the design and adoption of technologies to change and/or reinforce sustainable waste management behaviors across the globe (Suruliraj et al., 2020b). While various studies have emphasized that sustainable waste management apps contribute to promoting clean and sustainable environmental behaviors, however, they also reported a significant amount of disuse and abandonment (Comber et al., 2013). This is because, for behavior change to occur and for continued use of the sustainability apps, developers of the apps need to employ relevant persuasive strategies (Nkwo and Orji, 2018). These strategies give the app the ability to change, reinforce, motivate, and help users to adopt sustainable environmental behaviors that are potentially beneficial to them and their communities.

Previous research has conducted a literature review on the remote causes of inappropriate waste management (Omran and Gavrilescu, 2008; Ndubuisi-Okolo et al., 2016) or the design and evaluation of persuasive apps targeting specific waste management activities (Comber et al., 2013). However, to the best of our knowledge, no study has conducted a comparative systematic evaluation of sustainable waste management apps (on Google Play or App Store) across multiple sustainable waste management activities, using the behavior change strategies from the four categories of the persuasive system design (PSD) framework (Oinas-Kukkonen and Harjumaa, 2009).

To fill this gap, we conducted a comparative systematic evaluation of 148 apps that target various waste management activities. Some of the activities include personal tracking, recycling, conference management, data collection, food waste management, do-it-yourself (DIY) projects, games, etc. The goal of this evaluation is to identify and compare the persuasive strategies employed by the apps and how they were implemented. We coded the apps based on the persuasive strategies of the PSD framework. Although there are various persuasive principles, this study chose the PSD framework for its evaluation because it is more comprehensive and yield broader findings. Moreover, they have been used successfully in recent years to deconstruct and evaluate persuasive technologies to uncover strategies employed in motivating desirable behaviors among users in various domains such as health and wellness, physical activity, and environmental sustainability such as persuasive apps for waste management.

Among others, the findings from this study show that strategies from the primary task support (PTS) category were the most implemented in the apps, followed by system credibility support (SCS) strategies, dialogue support (DS) strategies, and social support (SS) strategies in descending order. Moreover, *reduction*, *personalization*, *real-world feel* and *surface credibility*, *reminder*, and *self-monitoring* were the most commonly employed persuasive strategies. In addition, there is a substantial relationship between the number of persuasive strategies employed and the apps' effectiveness as indicated by user ratings. Finally, we presented some design implications for tailoring such environmental sustainability apps to improve their effectiveness.

BACKGROUND AND RELATED LITERATURE

This section discusses literature associated with sustainable waste management. It defines the underlying principles and frameworks of persuasive designs. Also, it discusses relevant system development efforts and related literature that aimed to promote sustainable waste management activities and behaviors.

Sustainable Waste Management

Environmental sustainability is both a huge business and a global concern in line with the global climate change campaign. This is because sustainable waste management practices play a large and important role in guaranteeing the health and wellbeing of citizens and ensures a sustainable environment (Schiopu et al., 2007; Omran and Gavrilescu, 2008; Giusti, 2009). On the other hand, improper disposal of wastes is one of the leading causes of environmental pollution (Suruliraj et al., 2020a). Incidentally, the wastes can also be reduced, reused, and recycled to produce new and useful products, if properly managed (Abdul Rahman, 2000; Sridhar et al., 2014). Studies have shown that lack of awareness and negative attitudes are some of the hindrances to efficient waste disposal, sorting, and management in most developing communities (Nkwo, 2019). As a result, governments and stakeholders around the globe had put forward several

measures including awareness campaigns, legislation, and infrastructural supports, targeted at either motivating or compelling people to take on responsible waste management behaviors (Ndubuisi-Okolo et al., 2016; a et al., 2020). However, those efforts have not been effective, hence the calls for new approaches to motivate people to make behavioral and attitudinal changes. Such changes in behaviours can be realized through the combined powers of persuasion and emerging technologies. Specifically, this is when relevant persuasive strategies are implemented on user-centered technologies such as mobile phones (Nkwo, 2019).

Conventionally, persuasion involves "human communication intended to influence the autonomous judgments and actions of others" (Simons, 2011). The persuasiveness of technology is a function of its system qualities and techniques. Persuasive technologies (PTs) are interactive systems that utilize human–computer techniques or computer-mediated strategies. The strategies serve as building blocks of PTs, which are widely used in the environmental sustainability domain in general and sustainable waste management, in particular, to motivate and persuade users to change their attitudes, and support them to perform target behaviors.

Principles and Frameworks of Persuasion Design

Over the years, researchers have propounded several persuasion principles (Fogg, 2002; Cialdini, 2006; Fogg, 2009), frameworks (Oinas-Kukkonen and Harjumaa, 2009), and the goal-setting strategy (Locke and Latham, 2002), which could be employed to design, implement, and evaluate persuasive technologies. For instance, Fogg's functional triad and system design principles provided the original design concepts in persuasive technology development (Fogg, 2002). According to Fogg, three factors including motivation, ability, and triggers assist users to achieve their target behaviors. The main interest of Fogg's persuasion principle is to enhance these three factors to help researchers and designers to reflect more about the target behavior that needs to be promoted/reinforced or changed and understand how to design persuasive technologies to realize the objective (Fogg, 2009). However, certain weaknesses in the principles and theories such as "inability to translate design principles into actual software requirements" saw other researchers work to improve previous design recommendations to support design and evaluation activities.

Oinas-Kukkonen and Harjumaa, in their study, developed 28 design strategies based on three stages of PS development: 1) understanding the main issue behind PS, 2) analyzing the context of PS, and 3) describing different methods to design system features (Oinas-Kukkonen and Harjumaa, 2009). The strategies are referred to as the persuasive system design (PSD) framework and are classified into four distinct categories based on the type of support that the persuasive strategies provide to users of a system and application. These include the *primary task*, *dialogue*, *system credibility*, and *social support* categories (Nkwo et al., 2018; Oinas-Kukkonen and Harjumaa, 2009).

Table 1 shows the PSD framework categories, descriptions, and persuasive strategies. Also, **Table 2** shows a description of each of the strategies in the PSD framework.

In addition, the integration and operationalization of goal-setting strategy (a non-PSD strategy) into persuasive systems has been shown to increase task performance (Locke and Latham, 2002), directs people's attention, enhances their concentration, and lead to new approaches for performing target behaviors or tasks (van de Laar and van der Bijl, 2001).

Persuasive Strategies Employed in Designing Persuasive Apps for Waste Management

The PSD framework has been used to design persuasive technologies to promote sustainability behaviors. For example, Thieme et al. (2012) developed BinCam, which is a two-part design, combining a social persuasive system for the collection of waste-related behaviors (Thieme et al., 2012). BinCam is intended to blend seamlessly with the everyday routine of users, with the overreaching goal of making users reflect on food wastes and recycling behaviors of young adults, a playful and shared group activity. The findings from the evaluation of the intervention showed that users found the application interactive, supportive, socially collaborative, and effective in promoting food waste management and recycling behaviors. Subsequently, the BinCam social app was later redesigned and integrated with a Facebook app to improve engagement and motivate sustainable environmental behaviors (Comber et al., 2013). The findings from that study showed an increase in both users' awareness of, and reflection about, their waste management and their motivation to improve their waste-related skills (Thieme et al., 2012; Comber et al., 2013).

Another research carried out a user study of 153 students to discover factors that promote improper waste management behaviors among the students in a university campus in the global south. The findings from that study informed the design of a prototype waste management app, which could be used to encourage students to adopt clean and sustainable behaviors and protect the university environment *via* the provision of various personalized persuasive displays and support (Nkwo et al., 2018). The researchers employed relevant social influence strategies and personalization to tailor the design to the personal preferences and needs of the users, who were living in a closed community. Although the design was not evaluated, the results of that study demonstrated the potentials of using relevant persuasive strategies to encourage sustainable waste management behaviors among individuals and groups of people. It also showed how these strategies can be implemented on a computer and mobile technologies to help users to perform target behaviors without coercion. Subsequently, the researchers expanded their previous study to cover people living in a local community in South East Nigeria. The results of this study which were similar to the previous one were mapped to relevant persuasive strategies of the PSD framework. These strategies were used to develop socially appropriate design recommendations for building a mobile persuasive technology

TABLE 1 | PSD framework categories, descriptions, and persuasive strategies.

Category	Description	Persuasive strategies
Primary task support	Support users in performing their intended tasks	Reduction, tunneling, tailoring, personalization, self-monitoring, simulation, rehearsal
Dialogue support	Provide feedback that moves users toward the target behavior	Praise, rewards, reminders, suggestion, similarity, liking, social role
System credibility support	Support the development of more credible systems	Trustworthiness, expertise, surface credibility, real-world feel, authority, third-party endorsements, verifiability
Social support	Motivate users through social influence	Social learning, social comparison, normative influence, social facilitation, cooperation, competition, recognition

TABLE 2 | Description of each persuasive strategies of the PSD framework

Persuasive strategy	Description
Reduction	Reduces users' effort by breaking complex behaviors into simple to help them perform the target behavior
Tunneling	Guide users through a process to provide opportunities to encourage them along the way
Tailoring	Provide information that will be more persuasive if it is tailored to the potential needs, interests, personality, usage context, or other factors related to a particular user group
Personalization	Offer personalized content or customized services for users
Self-monitoring	Allow users to track and monitor their performance, progress, or status in achieving their goals
Simulation	Enable users to observe the link between the cause and effect of their behaviors
Rehearsal	Provide means for users to rehearse their target behavior
Praise	Offer praise through symbols, words, images, or sounds as feedback for users to encourage their progress toward the target behavior
Rewards	Provide virtual rewards for users when completing their target behaviors
Reminders	Remind users of their target behavior to assist achieve their goals
Suggestion	Provide appropriate suggestions for users to achieve their target behaviors
Similarity	Remind users of themselves or adopt trending features in a meaningful way
Liking	Contain a visually attractive look and feel which meets users' desires
Social role	Adopts a social role such as provide communication between users and the system's specialists
Trustworthiness	Provide truthful, reasonable, and unbiased information for users
Expertise	Provide information showing competence, experience, and knowledge
Surface credibility	Contain a competent look and feel that promote system credibility based on users' initial assessments
Real-world feel	Show information about people or organizations behind the content or services
Authority	Refer to people in the role of authority
Third-party endorsements	Highlight endorsements from respected and well-known sources
Verifiability	Provide means to investigate the accuracy of the content *via* external sources
Social learning	Allow users to observe other users' performance and outcomes while they are doing the same target behavior
Social comparison	Allow users to compare their performances with other users
Normative influence	Allow users to gather with other individuals who share the same objectives to feel norms
Social facilitation	Enable users to discern other users who perform the target behavior
Cooperation	Motivate users to cooperate with other users to achieve the target behavior goal
Competition	Motivate users to compete with other users to achieve the target behavior goal
Recognition	Provide public recognition, such as ranking feature, for users

to promote positive waste management behaviors among communities in the global south (Nkwo, 2019).

Persuasive Strategies Employed Based on Literature

Existing research has systematically evaluated mobile apps across several domains to establish the persuasive features they offer. For instance, in the health domain, researchers employed the strategies of the PSD framework to evaluate the effectiveness of web-based health interventions. The findings show that the intervention strategies especially the primary task support strategies were frequently implemented to encourage the adoption of healthy habits and behaviors (Kelders et al., 2012). Similarly, Orji and Moffatt (2016) conducted an empirical review of 85 papers to understand the effectiveness of persuasive technologies for health and wellness. The results of that study show that *self-monitoring*, which is one of the strategies in the primary task support category of the PSD framework, is most commonly used to operationalize persuasive health interventions (Orji and Moffatt, 2016). Based on these results, certain design recommendations were put forward to enhance the effectiveness of such health and wellness intervention. Furthermore, a systematic review of 32 papers was carried out to examine the effectiveness of social support strategies in encouraging physical activity. The results from that study show that *competition, social comparison*, and *cooperation*, which are among the strategies in the social support category of the PSD framework, were effective strategies used to motivate physical activity (Almutari and Orji, 2019). It recommended new approaches to tailor persuasive

interventions to support appropriate physical activities for various categories of users. In another study, 20 research papers that presented the design and evaluation of mobile apps for promoting physical activities were systematically evaluated (Matthews et al., 2016). The results of that study showed that although some other strategies such as *reduction*, *real-world feel*, and *personalization* were incorporated in the app design, *self-monitoring*, which is one of the strategies from the primary task support category, was the prevailing strategy employed in designing the apps. In addition, previous studies had uncovered that a g*oal-setting* strategy has the potential to increase task performance (Locke and Latham, 2002), direct people's attention, enhance their concentration, and lead to new approaches for performing target behaviors or tasks (van de Laar and van der Bijl, 2001). For instance, the results of a study that sought to suggest guidelines for designing persuasive apps to support improved breastfeeding behaviors show that such systems should allow users to set short, realistic, and measurable/trackable (self-monitoring), as well as incremental breastfeeding goals will lead to increased self-efficacy. The implication is that a relevant persuasive strategy from the PSD framework (Oinas-Kukkonen and Harjumaa, 2009) can be combined with the g*oal-setting* strategy to achieve a designed goal in a behavior-change intervention. This is important and offers great promises for designing user-centered software interventions aimed at promoting clean and healthy behaviors in the sustainability domain.

However, in the sustainable waste management sub-domain of the environmental sustainability domain, fewer recent studies have evaluated the persuasive strategies implemented in mobile apps for waste management. For instance, recent research was conducted to systematically review the persuasive strategies employed in the design of 125 sustainable waste management apps to identify the strategies from the primary task support category (alone) employed in app design (Suruliraj et al., 2020a). The results from that study showed that persuasive strategies such as *reduction*, *personalization*, *tailoring*, *self-monitoring*, and *rehearsal* were most commonly implemented in the apps in decreasing order. However, it also found no association between the number of persuasive strategies employed in the app's design and its effectiveness. This is in contrast to previous studies in other domains such as physical activity (Alhasani et al., 2020), where there was some level of relationship between the number of persuasive strategies employed in the app's design and its effectiveness. These findings draw attention to some huge gaps in research in this domain, which can be filled by a broader systematic evaluation of apps for sustainable waste management to uncover what persuasive strategies from the four categories of the PSD framework were employed in their designs.

Therefore, rather than evaluate apps to discover the persuasive strategies from the primary task support category alone, this current research article provides a comparative systematic evaluation of 148 apps across various sustainable waste management activities using the strategies from the four categories of the PSD framework (see **Table 1**). Specifically, we evaluated and compared the persuasive strategies from the primary task support, dialogue support, system credibility, and social support categories of the PSD framework and how they were implemented across the waste management activities such as personal tracking, recycling, conference management, data collection, food waste management, do-it-yourself (DIY) projects, and games, to uncover new insights and enrich the literature.

METHOD

This study aims to conduct a systematic review of sustainable waste management apps to identify and compare persuasive strategies (from the PSD framework) employed by the apps and how they were implemented to promote appropriate waste management behaviors. Therefore, we aim to address the following research questions:

1) What persuasive strategies were employed in designing the apps for sustainable waste management?
2) How were these strategies implemented on the apps to support targeted waste management activities?
3) Is there any relationship between the number of persuasive strategies employed in the app and the apps' effectiveness based on user ratings?

The answers to these research questions would help to inform our design recommendations for personalizing and tailoring sustainable waste management apps to improve their persuasiveness and effectiveness. The following subsections describe the apps' selection and filtering criteria and coding process.

Selection of Apps for Sustainable Waste Management

The app search for this study was carried out in 2020 during which we found out that most of the apps were updated last in 2019 (see **Supplementary Appendix** for details). We filtered our search results by selecting apps that matched with the following search terms: "waste management", "waste disposal", "waste recycling", "waste tracker", and "sustainable waste" on the App Store and Google Play. Second, we combined the search terms using "OR" and "AND" to search. The search results returned an initial list of 244 apps (App Store and Google Play).

We employed several criteria to extract the apps that best suit the objective of the study. Primarily, we accepted only those apps that are designed to support diverse waste management activities, are free or free with in-app purchases, are in English according to the app's description and demo, and have screenshots supplied in the description of every application. On the other hand, we excluded the apps that 1) do not support waste management activities, 2) were not described in the English language, 3) were not publicly available, 4) were outdated, and 5) cannot be logged in to explore its features and design strategies. Incidentally, the apps in this range had less than five ratings. Moreover, the researchers ensured that apps that appeared in both the App Store and Google Play were counted as one instead of two. In the end, a total of 148 apps were accepted and considered suitable for

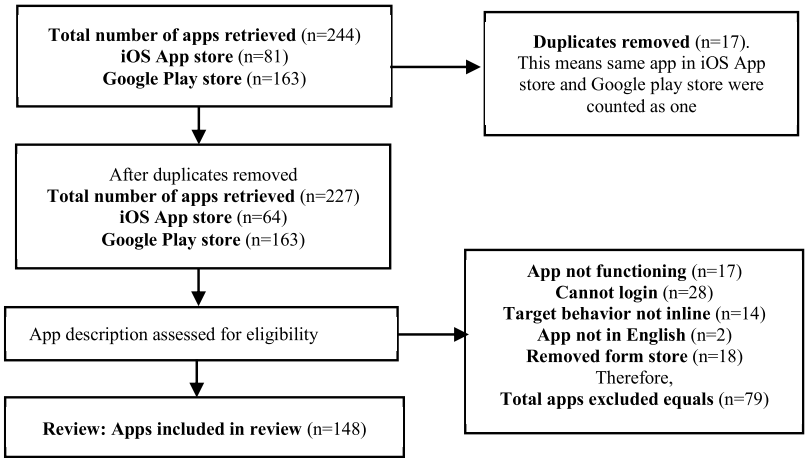

FIGURE 1 | Flow diagram illustrating selection and exclusion criteria in various stages.

coding (see **Figure 1** below). Some other information collected for each accepted app includes *application name, platform* (i.e., iPhone, Android, or both), *average rating*, *developer information*, *last update date*, and *price* (i.e., free, fee-based, and free with in-app purchases—where developers provide a free version and a paid version if users want to upgrade or unlock additional features in the app). Other information collected includes strategies implemented on the app, target outcomes, and country/region of development. We decided to choose the exclusion threshold of five ratings because it is the highest rating such apps could get from user reviews. While the apps with less than five ratings (n = 79) were excluded, apps left after exclusion were (n = 148). In other words, we selected 148 unique apps in total for coding and analysis. In addition, 85.6% of the apps were updated in 2019.

Process of Coding Apps for Persuasive Strategies

The purpose of coding the apps in our research is to evaluate the number and type of persuasive strategies employed in persuasive apps specifically related to sustainable waste management. Therefore, we identified the persuasive strategies (PSs) employed in designing each of the 148 sustainable waste management apps including how the strategies were implemented using the PSD framework. We chose this framework because it is more comprehensive and yields broader findings. It has been widely used in deconstructing and evaluating persuasive technologies across various domains. Two of the authors who are persuasive technology researchers installed the apps on their smartphones (Android and iOS) and used the app features to perform various tasks while taking note of the PSs integrated into them and how they were implemented, in our coding sheets. All the PSs were under the primary task support, dialogue support, system credibility support, and social support categories for coding purposes. The coding sheet was adapted from previous literature (Orji and Moffatt, 2016), validated by Nkwo et al. (2020), and modified for this research. For the features of the in-app purchase, researchers accepted the free trial to enable the examination of all persuasive strategies employed in the apps. The interrater agreement score for each strategy was computed afterward. Finally, a third expert reviewer was involved in resolving any disagreement for strategies having agreement less than 100%. **Figure 2** presents the steps of coding the apps. See **Supplementary Appendix** for the summary of the apps evaluated and the persuasive strategies employed by the apps.

Analysis of Data

We measured the percentage of agreement between two researchers (i.e., before the intervention of the third researcher—when needed). We also calculated interrater reliability using the percentage of agreement metric. Furthermore, we conducted descriptive statistics to obtain the average persuasive strategies employed in the design of the app. Finally, we examined the relationship between the number of persuasive strategies and the apps' effectiveness (based on the apps' ratings). Specifically, we performed a Pearson's correlation analysis between the number of persuasive strategies and the app's rating.

Computing correlation is important because it helps to explore the nature of the relationship between the two variables in question—determine which variables are most highly related to a particular outcome (Samuel and Ethelbert, 2015). Moreover, it provides the platform for regression to predict the values of the dependent variable based on the known relationship that exists between the independent variable and the dependent variable. In recent years, both the App Store and Play Store have placed a higher amount of importance on app ratings and reviews. This is because apps that have higher ratings and reviews rank high in search and have a better chance of being found and downloaded by potential users. Also, according to a recent report (Canstello, 2018), six of the most important metrics to measure apps' success are the number of users, active users, retention, cohort analysis, and lifetime value. These metrics predominantly inform user ratings and reviews and are pointers to how effective the apps are in helping users to perform and achieve set goals.

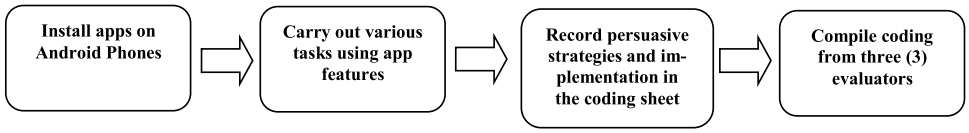

FIGURE 2 | PSD categories, descriptions, and their persuasive strategies.

TABLE 3 | Information on accepted apps

Mobile platforms	**iOS (23%), Android (77%)**
User ratings	5 (5.4%), 4–4.9 (57.4%), 3–3.9 (8%), 2–2.9 (2.6%), 1–1.9 (0.6), 0 or No rating (26%)
Waste management activity category	Productivity (21.6%), Education (15%), Business (15%), Lifestyle (13.5%), Food and Drink (9%), Social (4%), Other 15 categories (22%)

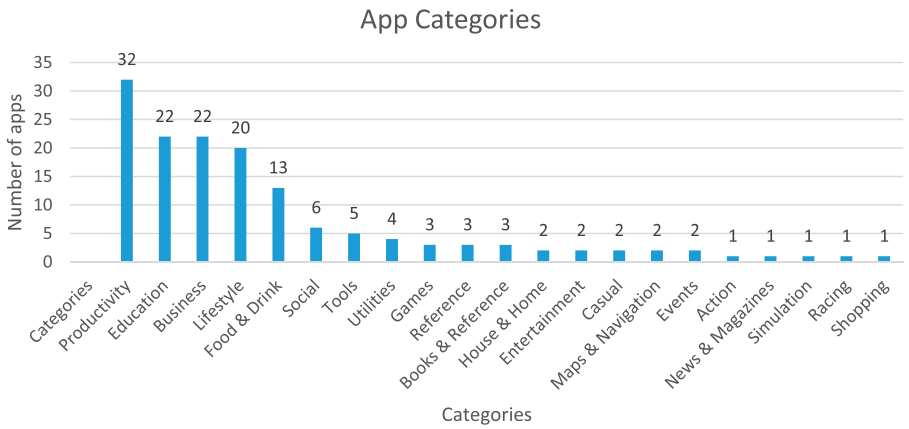

FIGURE 3 | Apps in each waste management activity.

Agreement

The interrater reliability for the coded apps was measured using the percentage of agreement metric as explained in Albert et al. (2017). Agreement occurs when the two reviewers both indicate the presence or absence of a persuasive strategy in an app. Disagreement occurs if one reviewer indicates the presence of a strategy, and the second reviewer indicates an absence. Reliability values range between 78.6 and 100% agreement depending on the persuasive strategy. The strategies with the lowest interrater reliability (78.6%) and (82.2%) were *normative influence* and *liking*, while 26 out of the 28 strategies obtained perfect agreement scores. Generally, all intercoder reliability scores were within the acceptable range (i.e., >60%) as described by Lombard et al. (2002).

RESULTS

This section presents the results of the study that provide answers to the three research questions itemized in the method section. Specifically, it discusses the persuasive strategies identified in the apps and how they were implemented across target sustainable waste management activities. It also discusses the relationship between the number of strategies employed and app effectiveness.

Information on Selected Apps

Table 3 shows the summary of the apps we downloaded and evaluated in this study. Sixty-eight percent (n = 100) of the apps were either released or updated in 2019. In addition, **Figure 3** shows the number of apps in each waste management category. Detailed information about the apps can be found in the **Supplementary Appendix**.

Persuasive Strategies Employed in Waste Management Apps

To answer research question 1, we downloaded 244 and evaluated 148 sustainable apps for waste management to uncover what persuasive strategies (from the PSD framework) were employed in their designs.

Generally, our findings show that 27 out of 28 different persuasive strategies of the PSD framework were employed in-app designs. We did not establish the implementation of the *social*

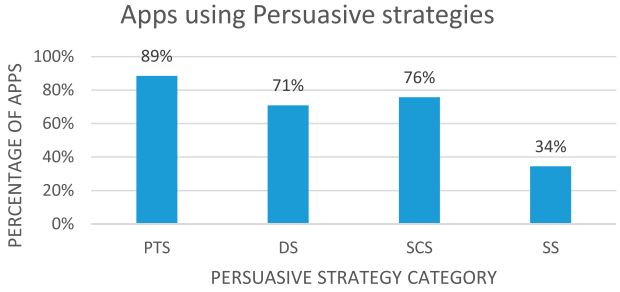

FIGURE 4 | Percentage of mobile apps employing persuasive strategies.

role strategy in any of the apps. The number of strategies employed in each app varied and ranges between 0 and 20. The hierarchical chart in **Figure 4** shows that the *primary task support* (PTS) strategies were employed the most 89% (n = 131), followed by the *system credibility support (SCS)* 76% (n = 112), *dialogue support* (DS) 71% (n = 105), and *social support* (SS) is least 34% (n = 51). We note that most of the apps employed more than one strategy in their implementations.

Also, the results from **Table 4** show that the strategies from the PTS category are the most employed in the sustainable waste management apps (sum = 327), followed by SCS (sum = 245), DS (sum = 190), and SS (sum = 75).

In addition, the persuasive strategies such as *reduction* (n = 97), *personalization* (n = 90), *self-monitoring* (n = 50), *real-world feel* and *surface credibility* (n = 83) each, and *reminder* (n = 73), *social facilitation* (n = 40) appear as the most frequently employed strategies in the reviewed apps. All other strategies were employed as follows: rewards (n = 36), suggestion (n = 33), verifiability (n = 32), praise (n = 29), liking (n = 18), rehearsal (n = 17), trustworthiness (n = 16), tunneling (n = 15), simulation (n = 13), expertise (n = 12), authority (n = 11), cooperation (n = 10), social comparison and third-party endorsement (n = 8 each), normative influence (n = 6), recognition and social learning (n = 4), competition (n = 3), and social role (n = 1). Please see **Figure 5** for a diagrammatic description of the strategies and corresponding number of apps implementing each of them.

Apps and Type of Waste Management Activities they were Designed for

To answer research question 2, we collected apps in 17 sub-categories based on the kind of waste management activities it was intended for (see **Table 3**). This was based on previous research (Suruliraj et al., 2020a). Among them, 34% (n = 51) apps were designed for regional waste disposal provided specifically to the local municipality. These apps primarily offer a garbage collection schedules calendar and waste sorting guide. Thirteen percent (n = 19) were designed to provide educational material such as articles, magazines, and news to educate people on waste management. Around 11% (n = 16) apps were focused to reduce food waste; apps in this category offer a marketplace for surplus food or track groceries in the refrigerator for expiry. Eight percent (n = 12) of the apps were used for commercial purposes and owned by private organizations. Commercial apps are used to request and manage on-demand services like dumpster rental in exchange for money. About 7% (n = 11) of the apps were designed as games; these apps will help the user to learn waste sorting by playing a sorting game and simultaneously provide facts. Some of the gaming apps offer points that can be redeemed for vouchers. In addition, 7% (n = 11) of the apps were developed for personal tracking. Personal tracking apps help users to track their daily waste management habits and show an impact chart for carbon emissions and plastics avoided. These apps can help to promote sustainable environmental behaviors. Six out of 17 sub-categories discussed previously cover 80% of the total apps evaluated in this study.

Figure 6 shows more information about other apps categorized according to their purpose and target behaviors.

In addition, **Figure 7** shows the persuasive strategies and types of waste management activities they were implemented for. Specifically, each of the waste management activities was operationalized with persuasive strategies as follows: Personal tracking and Conference (n = 9); Data collection, Food WM, and DIY projects (n = 7); Game, Cloth WM, and Regional waste disposal (n = 6); Marketplace and Calculator (n = 5); Magazine, Education, Plastic WM, and Commercial WM (n = 4); Biomedical WM and Waste collection (n = 3) and AI-aided waste sorting (n = 2).

Persuasive Strategies Implementation in the Apps

Generally, persuasive strategies are used to motivate and influence users to reach their personal and group goals through user engagement and collaboration. However, in this section, we present the distinct implementations of the strategies of the PSD framework, which are frequently employed in sustainable waste management apps.

TABLE 4 | PSD framework categories, persuasive strategies, and total strategies employed in apps

Category	Persuasive strategies	Total strategies in apps
PTS	Reduction (97), Tunneling (15), Tailoring (45), Personalization (90), Self-monitoring (50), Simulation (13), Rehearsal (17)	327
DS	Praise (29), Rewards (36), Reminders (73), Suggestion (33), Similarity (0), Liking (18), Social role (1)	190
SCS	Trustworthiness (16), Expertise (12), Surface Credibility (83), Real-world feel (83), Authority (11), Third-party Endorsements (8), Verifiability (32)	245
SS	Social learning (4), Social comparison (8), Normative influence (6), Social facilitation (40), Cooperation (10), Competition (3), Recognition (4)	75

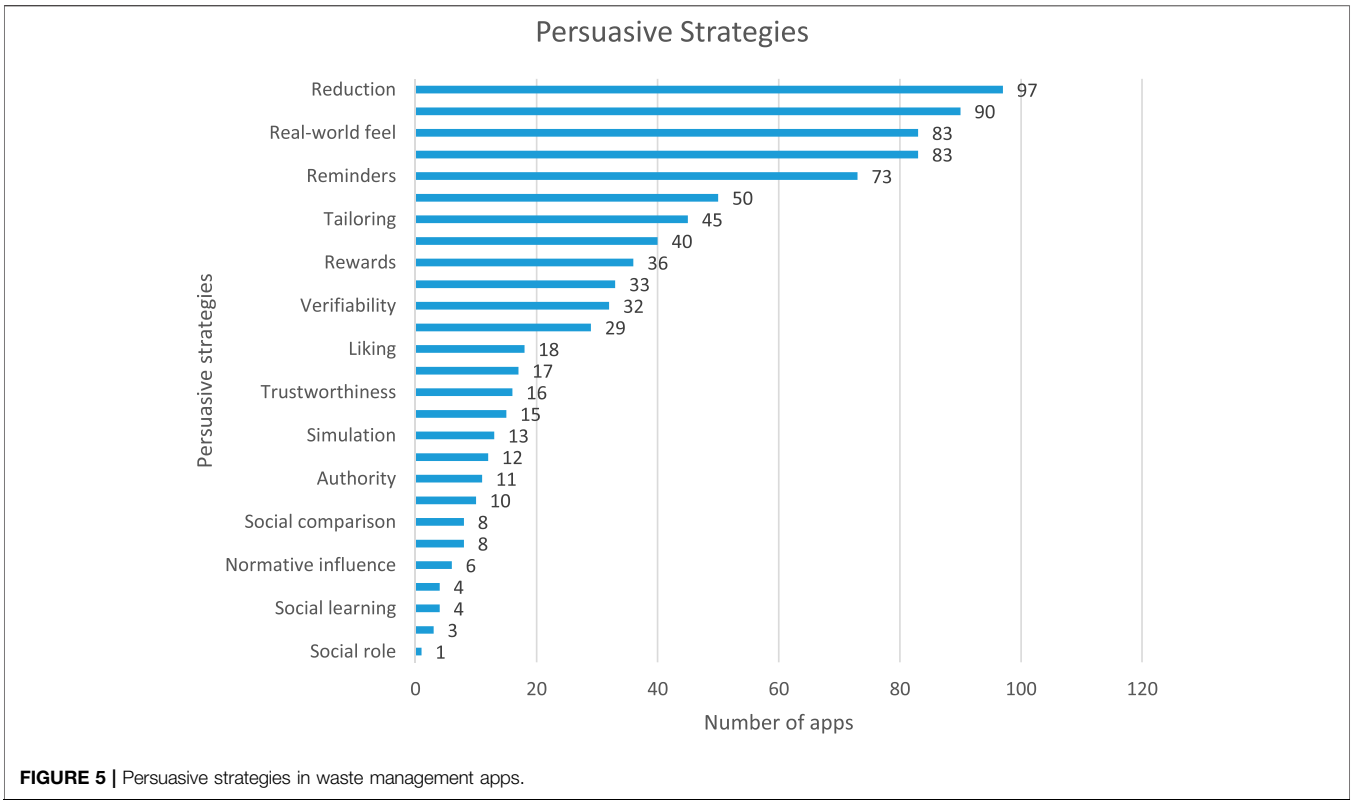

FIGURE 5 | Persuasive strategies in waste management apps.

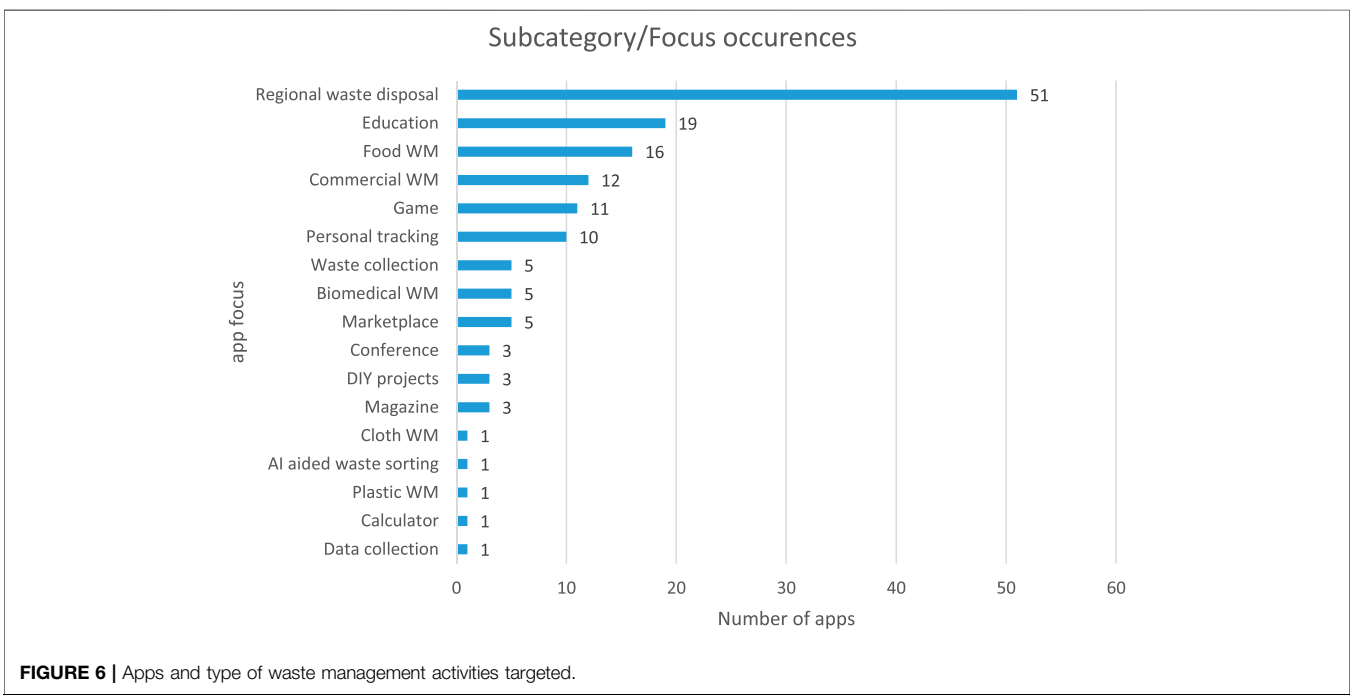

FIGURE 6 | Apps and type of waste management activities targeted.

Primary Task Support Strategies

The primary task support (PTS) strategies support individuals and groups to perform their primary tasks (Oinas-Kukkonen and Harjumaa, 2009). We found that 89% (n = 131) of the sustainable waste management apps implemented the strategies from the *primary task support* (PTS) category of the PSD framework (see **Figure 4**). The commonly implemented strategies in the PTS category are *reduction*, *personalization*, and *self-monitoring*

FIGURE 7 | Persuasive strategies and the type of waste management activity targeted.

among others (see **Figure 5**). Specifically, *reduction* strategies, which "reduce complex tasks into simpler ones so that system users can perform target behaviors easily" (Nkwo and Orji, 2018), were implemented in 97 apps as suggestive search (auto-populate listing) to reduce efforts in searching for relevant information. Other apps implemented it as a calendar view with color-coding to reduce time spent in searching for a garbage collection schedule, QR code/Bar code scan, and log in using third-party apps like Facebook and Google. *Personalization* strategies offer personalized content, functionalities, and services to users (Oinas-Kukkonen and Harjumaa, 2009), and were implemented in 90 apps as personalized language settings. These allowed users to choose the preferred languages with ease. Other apps implemented it through personalized notification times, email reminders, save location, user profiles, and personalized setting of user preferences and payment options. *Self-monitoring* strategies, which "allow people to keep track of their performances, offering information on both past and current behaviors" (Orji, 2017), were implemented in 50 apps as exclusive app screens to review trends of individual data related to history, statistics, environmental impact, and amount of CO_2 wastes released. The gaming apps implemented it *via* a real-time display of the player progress points earned and levels completed per game session.

System Credibility Support

The system credibility support (SCS) strategies describe how to design a system to be more credible and persuasive (Oinas-Kukkonen and Harjumaa, 2009). Seventy-six percent (n = 112) of the sustainable waste management apps implemented the strategies from the system credibility support (SCS) category of the PSD framework (see **Figure 4**). The commonly implemented strategies in the SCS category are *real-world feel* and *surface credibility* among others (see **Figure 5**). While *real-world feel* strategies provide information about the owners of the system, *surface credibility* strategies offer a competent look and feel for users (Nkwo and Orji, 2018). The real-world feel and surface credibility strategies were both implemented in 83 apps each through "about us/contact us pages", "terms of service", "privacy policy", version information with date, Frequently Asked Questions (FAQ) section, list of services offered, website information, and map view.

Dialogue Support Strategies

The dialogue support (DS) strategies offer some measure of system feedback to system users (Alqahtani et al., 2019). We uncovered that 71% (n = 105) of the sustainable waste management apps implemented the strategies from the DS category of the PSD framework (see **Figure 4**). The commonly implemented strategy in the DS category is *reminder* among others (see **Figure 5**). *Reminder* strategies allow a system to remind the user to perform target behaviors (Nkwo and Orji, 2018). They are implemented in 73 apps as push notifications to remind users about disposing of garbage, food item expiration alerts, news, and suggestions, etc. Other apps implemented it alongside self-monitoring strategies to remind users to track their data and status, and/or to perform certain waste management activities such as waste sorting, garbage collection, evacuation of waste bins *via* email reminders, text messages, pop-ups, and sounds.

Social Support

The social support (SS) strategies describe how to design a system to support users to perform target behaviors by leveraging social influence (Nkwo et al., 2020). We uncovered that 34% (n = 51) of the sustainable waste management apps implemented the strategies from the SS category of the PSD framework (see **Figure 4**). The frequently implemented strategy in this category is *social facilitation* among others (see **Figure 5**). *Social facilitation* strategy allows a system to offer a means to discern other individuals who are performing the target behavior (Nkwo and Orji, 2018). This strategy is implemented in 40 apps in the form of a community forum of users and regional waste managers. Connected users could see each other's activities, concerns, and suggestions or planned waste management activities. This will set the stage for users to exchange views or cooperate to tackle certain waste management issues and concerns *via* shared social communities such as a Facebook group for the app.

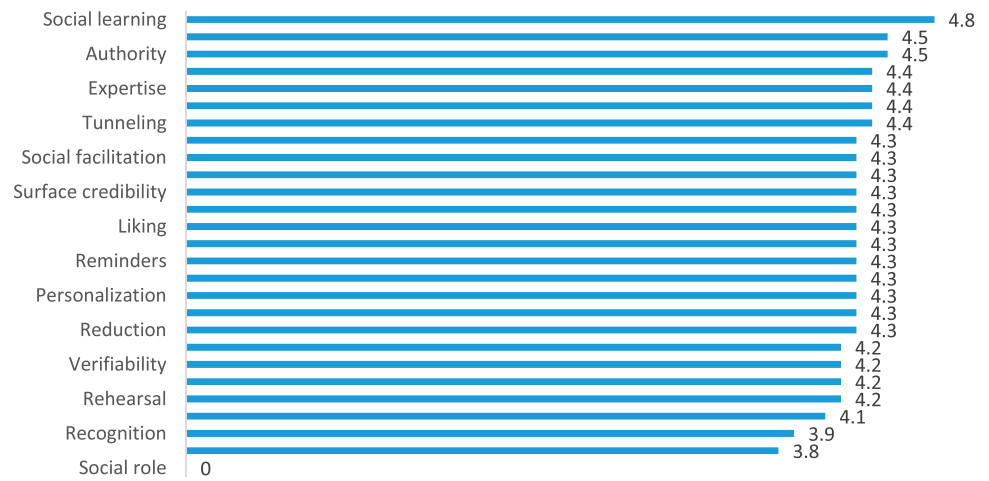

FIGURE 8 | Relationship between persuasive strategies and app ratings.

Persuasive Strategies and App Effectiveness

To answer research question 3, we ran the Pearson correlation coefficient (r) to determine whether any relationship exists between the number of persuasive strategies implemented in the apps and the apps' perceived effectiveness (based on average ratings). The computation was performed for all the apps combined. The results revealed that $r\,(146) = 0.21$, $p = 0.012$. The result means that overall, there is a significant correlation between the number of persuasive strategies employed and app effectiveness. This relationship that exists confirms the perceived effectiveness of the apps to promote sustainable waste management behaviors, from the user's point of view. Nevertheless, it is possible but unlikely that the correlation would change in this study's current state when different exclusion criteria and values are picked. This is because the exclusion criteria applied in filtering the apps with less than five ratings are fixed. In specific terms, we excluded the apps that did not support waste management activities, were not described in the English language, were not publicly available and those that cannot be logged in to explore its features and design strategies.

Furthermore, **Figure 8** shows that apps using the "*social learning*" strategy have the highest average rating of 4.8. All other strategies have their ratings as follows: "social comparison" and "authority" (4.5 each), "third-party endorsement", "expertise", "simulation" and "tunneling" (4.4 each), "cooperation", "social facilitation", "real-world feel", "surface credibility", "trustworthiness", "liking", "suggestion", "reminders", "self-monitoring", "personalization", "tailoring" and "reduction" (4.3 each), "normative influence", "verifiability", "rewards" and "rehearsal" (4.2 each), "praise" (4.1), except "*recognition*" and "*similarity*" strategies with 3.9 and 3.8, respectively. Only one app employed the "*social role*" strategy, but the app did not have any rating and was excluded. The average rating is a measure of what a given customer base or population, on average rates a certain product or service. It is computed using the following equation given the total number of ratings at each level.

$$AR = (1*a) + (2*b) + (3*c) + (4*d) + (5*e)/5$$

Where AR is the average rating, a is the number of 1-star ratings, b is the number of 2-star ratings, c is the number of 3-star ratings, d is the number of 4-star ratings, and e is the number of 5-star ratings.

DISCUSSION

In this section, we discuss the results of our study and offer some design recommendations for sustainable waste management apps based on our results and conceptual analysis as well as other relevant research.

Persuasive Strategies and Implementation

The goal of this research is to identify distinct persuasive strategies integrated into the apps developed to promote sustainable waste management behaviors and group the strategies based on the type of waste management issues or activities that the app targets or focused on. Furthermore, the study aims to uncover how the persuasive strategies were implemented in sustainable waste management apps to achieve their intended purposes, and also to examine the relationship between the persuasive strategies employed and apps' effectiveness.

First, this subsection provides answers to research question 1. It discusses the relevant persuasive strategies employed in designing the apps. Overall, the sustainable waste management apps reviewed in this paper employed 27 persuasive strategies.

The implementation ranges from minimum (0) to maximum (20) per app.

Primary Task Support Strategies

Predictably, we uncovered that the persuasive strategies from the primary task support (PTS) category of the PSD framework were the most employed in the apps 89% (n = 131). Among the strategies in this category, we discuss the implementation of three different strategies: *reduction, personalization*, and *self-monitoring*. We opted to discuss these three strategies because they are the most commonly employed strategies in the evaluated mobile apps. This is in agreement with a previous study (Suruliraj et al., 2020a), which shows that the primary task support strategies such as *reduction*, *personalization*, and *self-monitoring* among others are considered the main features of many sustainability interventions.

Reduction strategies emerged as the most implemented strategy (n = 97) and help users reduce efforts and simplify complex tasks into simpler ones so that users can be able to perform target behaviors with ease. The implementation of this strategy enables users to be able to search for relevant information such as the nearest public waste bucket, garbage collection schedules, etc. *via* a calendar view with color-coding. This reduces search time. System interventions that provide easier avenues to carry out target behaviors would motivate users to engage with and continue with the behaviors. These results demonstrate that the intervention strategies from the primary task support could be effective in helping individuals and groups to carry out their basic tasks or activities with ease. We refer to this attribute as "user-friendly routines". This finding is in agreement with previous studies (Oinas-Kukkonen and Harjumaa, 2009).

Personalization strategies emerged as the second most implemented persuasive strategy (n = 90) in sustainable waste management apps. Price et al. (2016) opine that allowing users to change colors, set backgrounds, and make other personalized settings on an app would improve its usability (Price et al., 2016). The ability to regulate the system intervention delivered *via* sustainable waste management apps to suit the user's needs and characteristics will make the system more effective. Moreover, studies have shown that personalized persuasive technologies are more effective at motivating users to perform target behaviors than the one-size-fits-all method of design (Moses et al., 2018). This is also true for sustainable waste management interventions in particular according to a recent study (Suruliraj et al., 2020a). So, it is unsurprising to see that sustainable waste management apps integrated some form of personalization because potential users may have unique needs and requirements based on factors such as literacy level, etc. This strategy will improve the user-friendliness of the app. We refer this attribute to as "adaptive design". This will allow users to customize certain functionalities of the app to improve its usefulness.

Self-monitoring is the third most employed strategy of the primary task support. It helps users of sustainable waste management apps to keep track and effectively manage their performances and goals (Matthews et al., 2016; Orji et al., 2018). Users can track their feeling, thoughts, and behaviors, which in turn increases self-awareness and motivate sustainable behavior outcomes. Most of the apps allowed for manual data entries and automatic display of information and statuses in the English language. Manual entries may be difficult and time consuming, and the display of user statuses in non-indigenous languages may not work for people with low literacy levels as they will not be able to read and write in English. The results demonstrate that many individuals or groups will be more motivated to embark on a task if they are provided with the means to keep track of their performance or status. We refer to this attribute as "performance tracking". Performance tracking is supported by intervention strategies such as self-monitoring, recognition, praise, and goal-setting. This finding is in line with previous studies (Orji et al., 2012, van de Laar and van der Bijl, 2001).

System Credibility Support Strategies

The persuasive strategies from the system credibility support (SCS) category of the PSD framework were the next most employed strategies in the apps 76% (n = 112). The credibility strategies such as *real-world feel* and *surface credibility* among others were implemented in sustainable waste management apps.

Real-world feel along with surface credibility emerged as the most implemented credibility strategy in the apps, and it provides information about people or organizations behind the app's content (Nkwo and Orji, 2018). It is offered in 83 apps. We argue that this strategy is essential in sustainability interventions. Like other interventions, apps for sustainable waste management should provide relevant and home-grown instructions, guidelines, and tips that are environmentally friendly and socially appropriate to users in a particular community. Anyone can design apps and publish them on the apps store, but technical and development skills are not sufficient for building apps that will effectively promote sustainable behaviors.

Surface credibility strategy is also offered in 83 apps. It ensures that the app offers a professional look and feel, to make a positive impression to users assessing the apps' contents and services (Nkwo and Orji, 2018). Considering that users will be supplying their sensitive information such as residential addresses, they need to be assured that their data are in credible hands. Full disclosure of owners' information and competent look and feel make an app credible (Oinas-Kukkonen and Harjumaa, 2009). Hence, providing opportunities for users to contact the app owners to make inquiries or ask questions and receive feedback from the apps, as well as ensuring a cleaner interface will improve the credit rating of an app.

Dialogue System Strategies

The persuasive strategies from the dialogue support (DS) category of the PSD framework were the third most employed strategy in the apps (71%, n = 105). The dialogue support strategy such as *reminders* among others was implemented in apps.

A reminder strategy is designed to remind users and improve their observance of desired behaviors. It reminds individuals about waste collection dates and locations, disposal of garbage, tracks their personal information, and to perform some helpful sustainable waste management activities such as sorting.

However, studies have shown that multiple and unsolicited reminders could annoy a user and lead to de-motivation and eventual disengagement (Bakker et al., 2016). There is therefore the need to take special cautions in implementing reminders in an app to avoid annoying users. One of the ways to achieve this result in an app is to tailor reminders to each individual or group. According to Alqahtani et al. (2019), tailoring reminders is significant because individuals and groups can be allowed to customize the frequency at which reminders are sent to them (how often), but also the type of reminder (pop-up boxes, text message, sounds, etc.) and when it should be sent (time). The results show that the strategies from the dialogue support could be useful in providing some degree of system feedback to its users, potentially through automated text messages, and pictorial or verbal information. We refer to this attribute as "automated notification management". This finding is in line with previous studies (Orji et al., 2012).

Social Support Strategies

The persuasive strategies from the social support (SS) category of the PSD framework were the fourth most employed strategies in the apps 34% (n = 51). Among other strategies in this category, *social facilitation* was the most implemented in the apps (see **Figure 5**).

Social facilitation is designed to provide a way to discern other individuals who are performing the target behaviors (Nkwo and Orji, 2018). It was implemented in 40 apps. Systems that offer opportunities for users to share their thoughts and concerns with similar others and build synergy with them will help to improve engagement. Users can share app-supplied information with other users *via* text, social media, email, or other means, depending on the device options. Therefore, developers of apps for a sustainable environment should focus on incorporating social facilitation features that allow users to recognize other users performing the same behaviors. This way the app will be more persuasive. Leveraging social influence strategies such as social facilitation could help shape users' behaviors. We refer to this attribute as "social support". This finding is in line with previous studies (Oinas-Kukkonen and Harjumaa, 2009).

Persuasive Strategies Implemented and Type of Waste Management Activities

Secondly, this subsection provides answers to research question 2. It discusses the type of sustainable waste management activities that the persuasive apps were designed for and how relevant persuasive strategies were implemented to support those activities. As can be seen from **Figures 6** and **7**, nearly all the sustainable waste management apps that we reviewed in this study targeted a mixture of waste management issues or activities. This makes it difficult to determine which persuasive strategies are more effective for a definite waste management activity. However, *reduction*, *personalization*, *self-monitoring* (primary task support), and *reminder (dialogue support)*, *real-world feel* and *surface credibility (system credibility support)*, and *social facilitation (social support)* are the most employed persuasive strategies in various sustainable waste management activities.

In general, the apps mostly targeted the following sustainable waste management issues or activities: personal tracking, conference management, data collection, food waste management, do-it-yourself (DIY) projects, games, and so on (see **Figure 7**). Specifically, apps for personal tracking and conference management employed the most number of strategies averaging nine strategies per app, followed by apps for data collection, food waste management, and DIY projects each with an average of seven strategies per app. Mobile apps that were designed as a game (waste sorting and recycling), cloth waste management, and regional waste disposal each implemented an average of six strategies. The marketplace and calculator apps employed an average of five strategies; apps focusing on the magazine, education, plastic waste management, and commercial waste management employed an average of four categories each. Mobile apps in the biomedical waste management and waste collection subcategories are second to the last, implementing an average of three strategies and artificial intelligence (AI)–aided waste management app implemented the least number of strategies; 2. For details, see **Figure 7**.

Persuasive Strategies Implemented and App Effectiveness

Thirdly, this subsection provides answers to research question 3. Specifically, the effectiveness of the apps was measured based on the app's rating. Interestingly, we established a significant relationship between the number of persuasive strategies and apps effectiveness as indicated by user ratings. This is particularly an interesting result considering the recent discussion and open research question on whether persuasive systems employing multiple persuasive strategies are more effective than those employing a single strategy (Orji, 2017). Our result implies that employing multiple strategies will increase apps' effectiveness in the area of waste management. This is not so with results from previous research in the health domain, which shows that employing one strategy can be effective (Alqahtani et al., 2019).

A possible explanation for the difference can be found in the differences inherent in the domains of investigation. This study targets sustainable waste management while the previous studies focused on health. For the previous study, it may seem that many people are conscious of their health since it has a personal and direct impact on their wellbeing—hence, they could easily be persuaded to adopt a healthy behavior. However, this is not the same with the sustainability domain (especially sustainable waste management), which has more of an indirect and most time community-level effect. It may take some extra effort to motivate people to adopt sustainable waste management behavior since it is difficult to show the cause-and-effect of each individual's behaviors and their contributions to the global, national, and community sustainable development goals (SDGs). Hence, designers and other stakeholders must focus on selecting the appropriate combination of persuasive strategies for an app, having both the target users and target activities in mind.

Comparative Evaluation of Dominant Persuasive Strategies

Table 3 describes the leading persuasive strategies employed in the apps. In a fast-paced world where ease of access and exactness are needed, *reduction* and *personalization* are certainly vital to tailor sustainability apps to individual users. It is therefore not surprising that reduction and personalization are the most dominant and most implemented in sustainable waste management apps. Users tend to be critical and may abandon apps if it is not user-friendly and does not support personalized access. While *reminders* and *suggestions* are important for notifying, reminding, and providing feedback to users to perform a target behavior, *praise* and *reward* are essential for providing positive reinforcements using virtual praise and/or rewards (e.g., texts or badges or sounds) or real rewards (e.g., coupons). These are important for the continued performance of target behaviors. *Self-monitoring* is also dominant in sustainable waste management apps since technological advancements have made it possible to automatically track personal and performance data over time, public trash can, etc., in real time through various sensors on smartphones, wearable devices, and public facilities. This will help users and managers to visualize their daily contributions to a clean and sustainable environment, and help them become more responsible and conscientious citizens of the society. It is also possible to monitor food wastes and carbon monoxide emission levels in industrial settings using tracked information. This explains why self-monitoring is among the top in the domain of environmental sustainability. *Surface credibility* and *real-world feel* are important for integrity, emotion, and positive feelings, due to the sensitive nature of these apps. Users tend to be skeptical and critical of apps in these areas and that makes it essential that the apps must be professional-looking, responsive, and with a visually appealing interface to be adopted. Any app that lacks these attributes may be deemed incredible. Hence, surface credibility is one of the popular strategies in the sustainability domain. Relevant social influence strategies such as *normative influence*, *social facilitation*, and *social role* are significant and useful in motivating individuals and groups of users to perform desirable waste management behaviors through positive peer pressure, evidence-based information displays, etc.

Design Implication

In this section and based on our findings, we offer design suggestions for tailoring sustainable waste management apps to improve their persuasiveness and effectiveness. In addition, we carefully integrated into our design recommendations some findings from relevant research (such as goal setting—a non-persuasive system design strategy), which will potentially strengthen some of the persuasive features of the app and hence improve its effectiveness (see **Table 5).**

1) User-friendly Routines: Accessibility and the ease of use of the various features of the app may have a significant influence on the user's behaviors toward task performance. Therefore, the designer should employ the *reduction* strategy in apps that target sustainable waste management to help users to perform their primary tasks with less difficulty and when required. Providing essential and easily accessible features such as shortcut menus, single-click or one-button press commands to commonly requested waste management issues such as collection and disposal locations and times, waste sorting, etc., would reduce complex behaviors for busy people and encourage them to imbibe appropriate waste management lifestyles even on the go behaviors (Nkwo and Orji, 2018). For example, the app may be customized to list the locations of nearby public waste bins in a community. This feature could be configured (using Google Maps) to automatically detect the user's current location and suggest the closest waste drop-off location, thereby helping users to preplan their routes to work/business and dispose of their wastes at the appropriate places. Moreover, because of the low literacy rate in certain communities, especially in the Global South, technical knowledge or extensive smartphone usage skills cannot be assumed for every user of such mobile apps. Therefore, designers should simplify the process by presenting the most frequently accessed features and easy-to-use features to the potential users of the apps, all advanced features that can be accessed by experienced users may require more steps to access them. This will help reduce the amount of effort and time that users spend trying to figure out how to use the mobile app to perform an activity and focus on the intended waste management activity.
2) Adaptive Features: Offering personalized content and features which will allow users to adapt some app functionalities to suit their individual preferences will go a long way to motivate the performance of target behaviors and may increase the apps' effectiveness (Nkwo and Orji, 2018). Adjusting app features such as the font size, type, and color of texts, background, layout, type of wastes you want to dispose of, waste management activities that users want to engage in, etc., based on user's data, would improve the usefulness of the sustainable waste management interventions. Moreover, given that many sustainable waste management apps target more than one waste management issue or activity, it becomes imperative that designers adapt the apps based on the type of waste management issues or activities that each experience. In addition, individuals who may be engaged in similar or same waste management activities may have unique needs that require personalized attention, hence emphasizing the need to personalize sustainable waste management apps to each need. Similar to system-controlled adaptation, designers can enable user-controlled adaption (customizations). This will allow users to adapt the features and functionalities of the applications to suit their needs. Research shows that both approaches to adaptation share common strengths of increasing users' perception of a system's relevance, usefulness, interactivity, ease of use, credibility, and trust, and also increases users' self-efficacy (Orji et al., 2017). However, there are notable differences between system- and user-controlled adaptation. User-controlled adaptation gives users a sense of *freedom*, *control*, and *personal touch* over the system, which in turn increases their commitment and hence systems effectiveness. System-controlled adaption reduces the app complexity (Orji et al., 2017). Therefore, we recommend that app designers can employ both, providing some adaptable features that users can control themselves, including background color, font, allowing app features to be enabled or disabled, and removing unnecessary categories that do not apply to their waste management needs.

TABLE 5 | Practical recommendations for design and associated persuasive strategies

Recommendations for design	Persuasive strategies
User-friendly routines	Reduction
Adaptive features	Personalization, tailoring
Automated notification management	Reminder, praise, reward, suggestion
Performance tracking	Self-monitoring, goal-setting, recognition, praise, reminder, suggestion
Credibility and responsiveness	Real-world feel, surface credibility
Social support design	Social facilitation, normative influence, social role

3) Automated Intelligent Notification Management: Providing intelligent reminders to notify the user to perform their target behaviors or keep track of certain waste management activities would help to motivate sustainable waste management behaviors and increase the apps' effectiveness (Oinas-Kukkonen and Harjumaa, 2009). For example, the designer can implement a feedback mechanism to remind the user to dispose of the right kind of waste at the right time, notify a user about a food's expiry, or exciting waste-for-cash offers in nearby waste collection locations. For mobile apps that support personal tracking of waste disposal habits, persuasive reminders that motivate/reinforce positive benefits and reward compliance can motivate users to continue with desirable waste management behaviors. This aligns with research that shows that positive reinforcement and gain-framed appeal are possible intervention strategies for strengthening people's behaviors (Orji, 2017). Positive reinforcement (Wilson, 2003) can be achieved by rewarding every sustainable waste management act ("*praise*" and "*rewards*" strategies) using virtual praise and/or rewards (e.g., texts or badges or sounds) or real rewards (e.g., coupons). On the other hand, gain-framed appeal refers to notifications that focus on the benefits of adhering to or performing a target (Wansink and Pope, 2015) (e.g., waste disposal, waste sorting) and can be facilitated using the *suggested* strategy. For example, gain-framed messages like "*By sorting your waste appropriately, you'll get a chance to earn some cash.*" can be sent at specified times to people motivated. Multiple and unsolicited reminders could annoy a user and lead to de-motivation and eventual disengagement (Bakker et al., 2016). To avoid this scenario, designers should tailor reminders to each individual or group. The act of tailoring reminders would allow app users to customize the frequency at which reminders are sent to them (how often), but also the type of reminder (pop-up boxes, text message, sounds, etc.) and when it should be sent (time).
4) Performance Tracking: The designers should employ a self-monitoring strategy in apps that target sustainable waste management activities to track their data and performance over time. Allowing individuals to track their performance and visualize their data (performance statuses) in attractive formats would offer the opportunity for self-awareness and evaluation, and help them to become more responsible in managing their wastes. For example, if a user is convinced that reducing his daily level of carbon-dioxide emission in the locality is beneficial, there is a possibility that he will continue to perform target behaviors. Also, performance tracking can be achieved *via* the design of mobile apps that tracks and updates the display of user contribution to a clean and sustainable environment by cutting down plastic use, reselling old electronics, up-cycling old items, etc. An impact chart with categories of waste will potentially help the user to visualize their progress which may engender self-efficacy. Some behavior data cannot be automatically monitored without users' involvement due to technology limitations. Therefore, for such behaviors, designers should provide some forms of praise and/or reward to users for tracking their behaviors each day. Performance tracking techniques have been used to support motivated people, especially those who are experienced in the potentialities of such interventions, to achieve target behaviors. However, according to previous studies, inexperienced users will likely be more demotivated in the process of using performance tracking interventions (Rapp and Cena, 2016). This is not unconnected to cumbersome tasks associated with personal information collection, nonfigurative visualizations, and the use of technology (Rapp and Cena, 2016). This will even be more evident in local communities in the Global South as such behavior-change apps would be deployed among potentially low-literate users who may be more disinclined to new technology adoption. Therefore, there is a need to employ complementary strategies that will take away the cumbersome tasks and expectations from users of the app. In addition, taking the job away from users and automating the collection of personal data and display of relevant information to users in visually attractive and descriptive formats would motivate the usage of such apps among less literate users. The *reduction*, *similarity*, and *liking* strategies could be employed to achieve this purpose. They should be integrated to reduce the number of efforts needed to perform target behaviors, and remind users about themselves and desired target behaviors in a visually attractive manner. Other corresponding persuasive strategies such as *reminders* and *suggestions* should also be operationalized on such apps to remind and help users to track and record their data. Self-efficacy can be enhanced through self-commitment by setting short-term goals (van de Laar and van der Bijl, 2001). The integration of the "*goal-setting*" strategy will motivate task performance, channel people's attention and focus on desired behaviors, enhance their awareness, and lead to new approaches for succeeding in the task (Locke and Latham, 2002, van de Laar and van der Bijl, 2001). The goal should be incremental (Orji, 2017); in other words, as an individual's confidence grows, the set goal could be reviewed upwards. Hence, sustainability interventions such as waste management games/apps should allow users to set short, realistic, and measurable (*self-monitoring*), as well as incremental sustainable waste management goals. This will lead to increased self-efficacy.
5) Credible and Responsive Design: The apps should be designed to provide potential users with relevant and home-grown sustainable

waste management instructions, guidelines, and tips that are socially appropriate to a particular community. Anyone can design apps and publish them on the apps store, but technical and development skills are not sufficient for building apps that will effectively promote sustainable behaviors. Thus, the app should offer waste management information that is endorsed by expert third parties. The users should also be able to verify the reliability of the information presented on the app. This will increase app reliability and encourage users to engage with the app. Moreover, surface credibility ensures that the app offers a professional look and feel, to make a positive impression to users assessing the apps' contents and services (Nkwo and Orji, 2018). Considering that users will be supplying their sensitive information such as residential addresses, they need to be assured that their data are in credible hands. Full disclosure of owners' information and competent look and feel make an app credible (Oinas-Kukkonen and Harjumaa, 2009). Hence, providing opportunities for users to contact the app owners to make inquiries or ask questions and receive feedback from the apps, as well as ensuring a cleaner interface will improve the credit rating of an app.

6) Social Support Design: Employing strategies that leverage social influence to design apps for sustainable waste management will provide users the opportunity to share their experiences and support one another to perform target behaviors. A user can be able to discern others who are engaged in similar waste management activities and would be motivated to share her experiences and concerns with them (social facilitation). They can also share the app contents on other media (SMS, WhatsApp, Facebook, etc.), which helps to spread the word and will help to bring like-minded people together. Using the "*normative influence*" strategy, positive peer pressure can be applied to enhance the possibility that an individual will adopt positive waste management behaviors. For instance, education mobile apps that offer evidence-based sustainable waste management information and community resources (including inspiring photos/videos, success stories, testimonials, etc.) for educating and influencing changes in beliefs, narratives, or attitudes can be disseminated to target groups. This could be done through discussion forums (peer-to-peer, stage-matched, or moderated peer-to-peer forums), online mutual-help support communities, asynchronous bulletin boards, and virtual chat rooms. In addition, the "social role" strategy through the ask-a-waste-manager service can help to support individuals toward sustainable waste management.

LIMITATIONS

This study has several limitations. One of them is that we reviewed only apps that were provided in the English language. Since there are apps that are in other languages, the results may not be generalizable. Second, due to the dynamic nature of the Google Play and iOS App stores, the composition and features of the apps we reviewed could be altered by the time this paper is published. In addition, user ratings may not be enough to ascertain the effectiveness of apps. This is because many other factors can influence the effectiveness of apps. However, user rating was the singular, closest evaluation we had to measure effectiveness.

CONCLUSION

Our society has become a platformized one. Mobile technology, which is one of the major features of our society, is a major influencer and could be employed to promote sustainable behavior change. This article provides a systematic evaluation of mobile apps for sustainable waste management to deconstruct and compare the persuasive strategies employed and their implementations.

The results from this study show that strategies from the primary task support, followed by system credibility support, dialogue support, and social support categories, were implemented at various levels. Specific persuasive strategies such as *reduction*, *personalization*, *real-world feel* and *surface credibility*, *reminder*, and *self-monitoring* were regularly used to design the apps for sustainable waste management such that it could motivate users to perform target behaviors. Moreover, it discovered that there is a relationship between the number of persuasive strategies employed and the effectiveness of the apps. Lastly, based on the results, we presented design implications for tailoring such persuasive apps for sustainable waste management to improve their effectiveness. In future research, experimental work will be required to show the guideline's applicability in the actual design and usage situation of persuasive technologies for sustainable waste management in particular and environmental sustainability in general. Future studies will also examine which persuasive strategies are most important to users in achieving sustainable waste management goals.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

MN conceptualized the study. MN and BS collected and analyzed the data. All authors wrote the article. RO and MN reviewed the article. RO supervised the study.

ACKNOWLEDGMENTS

The researchers wish to thank the reviewers of this article for their comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.748454/full#supplementary-material

REFERENCES

a, O., Nkiruka, C., b, E., c, A., and D.A, B. (2020). Waste Management Policy Implementation in Nigeria: A Study of Rivers State Waste Management Agency. *Int. J. Adv. Res.* 8 (02), 755–765. doi:10.21474/ijar01/10506

Abdul Rahman, F. (2000). *Reduce, Reuse, Recycle : Alternatives for Waste Management*. New Mexico state: NM State University, 1–4.

Albert, M. A., Durazo, E. M., Slopen, N., Zaslavsky, A. M., Buring, J. E., Silva, T., et al. (2017). Cumulative Psychological Stress and Cardiovascular Disease Risk in Middle Aged and Older Women: Rationale, Design, and Baseline Characteristics. *Am. Heart J.* 192, 1–12. doi:10.1016/j.ahj.2017.06.012

Alhasani, M., Mulchandani, D., Oyebode, O., and Orji, R. (2020). "A Systematic Review of Persuasive Strategies in Stress Management Apps," in Proceedings of the Eighth International Workshop on Behavior Change Support Systems (BCSS2020) (Aalborg, Denmark: Conference: Persuasive 2020 - Behavior Change Support Systems).

Almutari, N., and Orji, R. (2019). "How Effective Are Social Influence Strategies in Persuasive Apps for Promoting Physical Activity? A Systematic Review," in ACM UMAP 2019 Adjunct - Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (Windhoek, Namibia: Association for Computing Machinery, Inc), 167–172.

Alqahtani, F., Al Khalifah, G., Oyebode, O., and Orji, R. (2019). Apps for Mental Health: An Evaluation of Behavior Change Strategies and Recommendations for Future Development. *Front. Artif. Intell.* 2, 30. doi:10.3389/frai.2019.00030

Bakker, D., Kazantzis, N., Rickwood, D., and Rickard, N. (2016). Mental Health Smartphone Apps: Review and Evidence-Based Recommendations for Future Developments. *JMIR Ment. Health.* 3, e7. doi:10.2196/mental.4984

Canstello, D. (2018). 6 Essential mobile App Metrics to Measure success. Available at: https://supermetrics.com/blog/mobile-app-metrics (Accessed September 14, 2021).

Cialdini, R. B. (2006). *Influence: The Psychology of Persuasion*. New York, NY: HarperCollins.

Comber, R., Thieme, A., Rafiev, A., Taylor, N., Krämer, N., and Olivier, P. (2013). "BinCam: Designing for Engagement with Facebook for Behavior Change," in BinCam : Designing for Engagement with Facebook for Behavior Change. INTERACT 2013, At Cape Town, South Africa, 99–115. doi:10.1007/978-3-642-40480-1_7

Fogg, B. J. (2002). *Persuasive Technology: Using Computers to Change what We Think and Do*. Los Altos: Morgan Kaufmann.

Fogg, B. J. (2009). *A Behavior Model for Persuasive designPersuasive Technology, Fourth International Conference, PERSUASIVE 2009*. California, USA: Claremont.

Giusti, L. (2009). A Review of Waste Management Practices and Their Impact on Human Health. *Waste Management.* 29, 2227–2239. doi:10.1016/j.wasman.2009.03.028

Gu, T. (2019). Insights into the World's 3.2 Billion Smartphone Users, the Devices They Use & the Mobile Games They Play. Available at: https://newzoo.com/insights/articles/newzoos-globalmobile-%0Amarket-report-insights-into-the-worlds-3-2-billion-smartphone-users-the-devicesthey-%0Ause-the-mobile-games-they-play/.

Gustafsson, A. (2010). Positive Persuasion Designing Enjoyable Energy Feedback Experiences in the home. Available at: www.ait.gu.se.

Istudor, I., and Gheorghe Filip, F. (2014). *The Innovator Role of Technologies in Waste Management towards Sustainable Development1st International Conference Economic Scientific Research - Theoretical, Empirical and Practical Approaches'*. Dambovita, Romania: ESPERA.

Kelders, S. M., Kok, R. N., Ossebaard, H. C., and Van Gemert-Pijnen, J. E. (2012). Persuasive System Design Does Matter: a Systematic Review of Adherence to Web-Based Interventions. *J. Med. Internet Res.* 14 (6), e152. doi:10.2196/jmir.2104

Locke, E. A., and Latham, G. P. (2002). Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-year Odyssey. *Am. Psychol.* 57 (9), 705–717. doi:10.1037/0003-066x.57.9.705

Lombard, M., Snyder-Duch, J., and Campanella Bracken, C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Hum. Commun. Res.* 28, 4587–4604. doi:10.1111/j.1468-2958.2002.tb00826.x

Matthews, J., Win, K. T., Oinas-Kukkonen, H., and Freeman, M. (2016). Persuasive Technology in Mobile Applications Promoting Physical Activity: a Systematic Review. *J. Med. Syst.* 40 (3), 72–13. doi:10.1007/s10916-015-0425-x

Moses, N., Lowens, B. M., Knijnenburg, B. P., Orji, R., and Sekou, R. L. (2018). *Cross-Cultural Perspectives on eHealth Privacy in Africathe Second African Conference for Human-Computer Interaction*. Windhoek, Namibia: Association for Computing Machinery.

Ndubuisi-Okolo, P., Rita Anekwe, I., and Yusuf Attah, E. (2016). Waste Management and Sustainable Development in Nigeria: A Study of Anambra State Waste Management Agency. *Eur. J. Business Management.* 8 (No.17).

Nkwo, M. (2019). "Mobile Persuasive Technology: Promoting Positive Waste Management Behaviors in Developing African Nations," in Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery.

Nkwo, M., and Orji, R. (2018). "Persuasive Technology in African Context: Deconstructing Persuasive Techniques in an African Online Marketplace," in Proceedings of 2nd African Computer-Human Interaction Conference (AfriCHI'18) Windhoek, Namibia, 10.

Nkwo, M., Orji, R., and Ugah, J. (2018). "Persuasion for Promoting a Clean and Sustainable Environment," in ACM International Conference Proceeding Series, Association for Computing Machinery, 259–262.

Nkwo, M., Suruliraj, B., Orji, R., and Ugah, Jn. (2020). "Socially-oriented Persuasive Strategies and Sustainable Behavior Change : Implications for Designing for Environmental Sustainability," in Persuasive 2020, Adjunct proceedings of the 15th International Conference on Persuasive Technology, 1–5.

Oinas-Kukkonen, H., and Harjumaa, M. (2009). Persuasive Systems Design: Key Issues, Process Model, and System Features. *Commun. Assoc. Inf. Syst.* 24 (1), 485–500. doi:10.17705/1cais.02428

Omran, A., and Gavrilescu, M. (2008). Municipal Solid Waste Management in Developing Countries: a Perspective on Vietnam. *Environ. Eng. Manag. J.* 7 (4), 469–478. doi:10.30638/eemj.2008.070

Orji, F., Vassileva, J., and Greer, J. (2018). "Personalized Persuasion for Promoting Students' Engagement and Learning," in CEUR Workshop Proceedings.

Orji, R., and Moffatt, K. (2016). Persuasive Technology for Health and Wellness: State-Of-The-Art and Emerging Trends. *Health Inform. J.* 24 (1), 66–91. doi:10.1177/1460458216650979

Orji, R., Oyibo, K., and F Tondello, G. (2017). "A Comparison of System-Controlled and User-Controlled Personalization Approaches," in UMAP 2017 - Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (Bratislava), 413–418. doi:10.1145/3099023.3099116

Orji, R., Vassileva, J., and Mandryk, R. (2012). Towards an Effective Health Interventions Design: An Extension of the Health Belief Model. *Online J. Public Health Inform.* 4, 3. doi:10.5210/ojphi.v4i3.4321

Orji, R. (2017). "Why Are Persuasive Strategies Effective? Exploring the Strengths and Weaknesses of Socially-Oriented Persuasive Strategies," in International Conference on Persuasive Technology (Cham: Springer), 253–266. doi:10.1007/978-3-319-55134-0_20

Paay, J., Kjeldskov, J., Skov, M., Pathmanathan, R., and Pearce, J. (2013). "Promoting Pro-environmental Behavior: A Tale of Two Systems," in Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, OzCHI 2013 (Association for Computing Machinery), 235–244.

Price, M., Sawyer, T., Harris, M., and Skalka, C. (2016). Usability Evaluation of a mobile Monitoring System to Assess Symptoms after a Traumatic Injury: a Mixed-Methods Study. *JMIR Ment. Health.* 3 (1), e3. doi:10.2196/mental.5023

Rapp, A., and Cena, F. (2016). Personal Informatics for Everyday Life: How Users Without Prior Self-Tracking Experience Engage With Personal Data. *Int. J. Human-Computer Stud.* 94, 1–17. doi:10.1016/j.ijhcs.2016.05.006

Samuel, M., and Ethelbert, O. L. (2015). The Relevance and Significance of Correlation in Social Science Research. *Int. J. Sociol. Anthropol. Res.* 1 (3), 22–28.

Schiopu, A-M., Apostol, I., Hodoreanu, M., and Gavrilescu, M. (2007). Solid Waste in Romania: Management, Treatment, and Pollution Prevention Practices. *Environ. Eng. Management J.* 6 (5), 451–465.

Simons, H. (2011). *Persuasion in Society*. thousand Oaks London New Delhi: Sage Publications, Inc.

Sridhar, M. K. C., Hammed, T. B., and Babatunde Ammed, T. (2014). Turning Waste to Wealth in Nigeria: An Overview. *J. Hum. Ecol.* 46 (2), 195–203. doi:10.1080/09709274.2014.11906720

Suruliraj, B., Nkwo, M., and Orji, R. (2020a). *Persuasive Mobile Apps for Sustainable Waste Management : A Systematic Review*. Switzerland: Springer International Publishing.

Suruliraj, B., Olagunju, T., Nkwo, M., and Orji, R. (2020b). "Bota : A Personalized Persuasive Mobile App for Sustainable Waste Management," in Persuasive Technology Conference Aalborg, Denmark, 1–14.

Thieme, A., Comber, R., and Julia Miebach (2012). *We've Bin Watching You : Designing for Reflection and Social Persuasion to Promote Sustainable Lifestyles*. Austin, Texas, USA: CHI.

van de Laar, K. E., and van der Bijl, J. J. (2001). Strategies Enhancing Self-Efficacy in Diabetes Education: a Review. *Sch. Inq. Nurs. Pract.* 15 (3), 235–248.

Wansink, B., and Pope, L. (2015). When Do Gain-Framed Health Messages Work Better Than Fear Appeals? *Nutr. Rev.* 73 (1), 4–11. doi:10.1093/nutrit/nuu010

Wilson, E. V. (2003). Perceived Effectiveness of Interpersonal Persuasion Strategies in Computer-Mediated Communication. *Comput. Hum. Behav.* 19 (5), 537–552. doi:10.1016/s0747-5632(03)00006-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Artificial Intelligence

SYSTEMATIC REVIEW
published: 20 June 2022
doi: 10.3389/frai.2022.862997

The Impact of Pedagogical Agents' Gender on Academic Learning: A Systematic Review

Marjorie Armando[1,2,3], *Magalie Ochs*[1*] *and Isabelle Régner*[2*]

[1] Aix Marseille Univ, CNRS, LIS UMR 7020, Marseille, France, [2] Aix Marseille Univ, CNRS, LPC, Marseille, France, [3] Pôle pilote Ampiric, Institut National Supérieur du Professorat et de l'Éducation, Aix-Marseille Université, Marseille, France

Virtual learning environments often use virtual characters to facilitate and improve the learning process. These characters, known as pedagogical agents, can take on different roles, such as tutors or companions. Research has highlighted the importance of various characteristics of virtual agents, including their voice or non-verbal behaviors. Little attention has been paid to the gender-specific design of pedagogical agents, although gender has an important influence on the educational process. In this article, we perform an extensive review of the literature regarding the impact of the gender of pedagogical agents on academic outcomes. Based on a detailed review of 59 articles, we analyze the influence of pedagogical agents' gender on students' academic self-evaluations and achievements to answer the following questions: (1) Do students perceive virtual agents differently depending on their own gender and the gender of the agent? (2) Does the gender of pedagogical agents influence students' academic performance and self-evaluations? (3) Are there tasks or academic situations to which a male virtual agent is better suited than a female virtual agent, and vice versa, according to empirical evidence? (4) How do a virtual agent's pedagogical roles impact these results? (5) How do a virtual agent's appearance and interactive capacities impact these results? (6) Are androgynous virtual agents a potential solution to combatting gender stereotypes? This review provides important insight to researchers on how to approach gender when designing pedagogical agents in virtual learning environments.

Keywords: virtual agent, gender, pedagogical agent, learning environment, gender stereotypes, systematic review

OPEN ACCESS

Edited by:
Elaine Harada Teixeira de Oliveira, Federal University of Amazonas, Brazil

Reviewed by:
Kaoru Sumi, Future University Hakodate, Japan
Carol M. Forsyth, Educational Testing Service, United States

***Correspondence:**
Magalie Ochs
magalie.ochs@lis-lab.fr
Isabelle Régner
isabelle.regner@univ-amu.fr

Specialty section:
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

Received: *26 January 2022*
Accepted: *19 May 2022*
Published: *20 June 2022*

Citation:
Armando M, Ochs M and Régner I (2022) The Impact of Pedagogical Agents' Gender on Academic Learning: A Systematic Review. Front. Artif. Intell. 5:862997. doi: 10.3389/frai.2022.862997

1. INTRODUCTION

Pedagogical agents are virtual characters in digital environments used to improve learning in educational settings (Mohtadi et al., 2014; Schroeder et al., 2017). They can take on different roles, such as expert, mentor, or motivator (Baylor and Kim, 2005). As shown in a meta-analytic review of 43 studies by Schroeder et al. (2013), pedagogical agents can have a positive effect on students' free recall ability, knowledge retention, and transfer of prior knowledge to new situations or problems. However, some characteristics of pedagogical agents may impact the learning process: for instance, how realistic the virtual agents' appearance is Baylor and Kim (2004), the way they communicate with learners, verbally or nonverbally, positively or negatively (Gratch et al., 2007; Pecune et al., 2016), or the way they deliver feedback, using voice, text, or both (Kim and Baylor, 2016).

Virtual agents' gender is another feature that users can perceive from the agents' appearance (Lee, 2003). Yet few studies have evaluated the impact of pedagogical agents' gender, which is surprising considering the amount of research in Social Cognition documenting the impact of the gender of both learners and teachers on academic learning. Social Cognition and human-to-human studies are particularly interesting in the domain of virtual agents, as research shows that individuals have a propensity to interact with virtual agents as if they were human (Nass and Moon, 2000). Research in Social Cognition and Cognitive Psychology can, therefore, be enlightening for understanding users' perception of virtual characters and the effect of these perceptions on their performance. This is why we present some major Social Cognition research on the impact of learners and teachers' gender on learners' academic outcomes. For instance, Sansone (2019) conducted a survey on the link between high school students' beliefs about women's abilities in math and science and their teacher's gender, finding that students were less likely to report that men are better than women in math/science when assigned to female teachers. Teachers' behavior can also impact girls' and boys' learning differently: a large scale survey conducted by Forgasz and Leder (1996) showed that students who perceived their math teachers to be interested in them as individuals were more likely to have functional beliefs about themselves in mathematics, and this was more critical for female learners than male learners. Core beliefs represent general and strongly held views about ourselves, others, and the world; they influence the way we react in different circumstances. Functional beliefs are rational thought patterns that are generally useful for individuals to achieve their goals (Ellis, 1962). In the forementioned study, math teachers' behaviors seemed to favor boys over girls: boys had more interactions with their teachers, teachers were more tolerant of boys' misbehavior, and they had higher expectations of boys (Forgasz and Leder, 1996). A meta-analysis conducted by Lindberg et al. (2010) from 242 studies published between 1990 and 2007 indicated that while male and female learners performed similarly in mathematics, female students reported higher anxiety, more discomfort, and lower interest and self-efficacy in math classes than male students. Parents themselves tend to attribute different explanations for their children's academic performance depending on their gender: they explain their sons' mathematical success as due to their natural talent, whereas they explain their daughters' as due to their effort (Yee and Eccles, 1988). These results were replicated by Räty et al. (2002) who also found that parents of boys evaluated their child's mathematical competence as higher than parents of girls, and parents of girls perceived them as surpassing boys in reading. Despite this, parents still attributed competence in reading as resulting from the effort of girls but to the natural talent of boys. By explaining their daughters' success in math as due to effort, the authors suggested that parents may undermine both their own and their daughters' estimation of their daughters' success in mathematics, hence raising possible doubts about their future success in a domain that they think gets increasingly complicated; meanwhile, they may encourage boys to develop greater confidence in their future success (Yee and Eccles, 1988).

All these differences reflect the influence of gender stereotypes that lead people to consider men to be better at math than women, and women to be better in liberal arts -such as literature, e.g.,- than men. In addition, studies have shown that the fear of being negatively stereotyped in a skill area produces negative thoughts, which in turn reduce individuals' working memory capacity and impair learning and performance (Schmader and Johns, 2003). This phenomenon, called *Stereotype Threat* (Steele and Aronson, 1995), applies to different stereotypes and social groups, such as boys in reading tests (Pansu et al., 2016) and girls and women in math tests (Régner et al., 2014). The effects of Stereotype Threat can be reduced using different strategies, such as reading a story about a successful role model before taking a test (Bagès and Martinot, 2011; Bagès et al., 2016).

Presenting pedagogical agents as role models could be a potential solution for reducing the effects of Stereotype Threat. Researchers designing agents should take into account the gender of both learners and pedagogical agents to adapt the agent to the learners. The advantages of adapting virtual agents to participants have been demonstrated in several studies. For instance, in Vilaro et al. (2021), participants (all Black women) liked Black female agents for being artificial, hence creating a sense of trust and freedom where participants could avoid inherent biases and racism. In virtual learning environments, research has shown the impact of virtual agents' gender on human-agent interactions (refer to Section 3.4). However, the gender of *pedagogical* virtual agents is rarely considered an important characteristic in the design of virtual learning environments, whereas most pedagogical agents are human-like, and their gender can have an impact on academic outcomes (refer to Section 3.5). In terms of perception, various studies have shown that male virtual agents are rated as more powerful (Nunamaker et al., 2011), more expert (Nunamaker et al., 2011), and more knowledgeable (Baylor and Kim, 2004), whereas female agents are rated as more likable (Nunamaker et al., 2011) and more attractive (Lunardo et al., 2016). These attributes are important in learning environments, as competent and expert agents improve learners' performance (Baylor and Kim, 2004), and likable and attractive agents improve learners' self-perception including their self-efficacy (feeling of achievement) (Rosenberg-Kima et al., 2008), which may help improve their performance (Plant et al., 2009).

In this article, we present an extensive state of the art focusing on the *effects of pedagogical agents' gender in virtual learning environments*. We explore the impact of gender on the users' perceptions of agents and on their learning.

This article is organized as follows. In the next section, we explain our methodology used to conduct the state of the art and particularly how we used the PRISMA method to select relevant articles (Webster and Watson, 2002). In Section 3, the selected articles are summarized in **Tables 1**, **2** to provide a comprehensive review of research on the impact of pedagogical agents' gender on learners' performance and self-perception in academic domains. We discuss the articles summarized in the

TABLE 1 | Summary of articles on perceptive studies of virtual agents depending on their gender, regardless of the application domain.

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Lee (2003)	• 1 MA • 1 FA • 2-D • Text • Cartoon • Adviser	• 28 MP • 88 FP • avg age N/A	Playing a multiple-choice game with an agent. Participants could change their answer after they were told the agent's answer. It was specified that the agent's answer might not be correct.	• Masculinity • Attractiveness • Competence • Trustworthiness • Persuasiveness (sport or fashion questions)	• *Masculinity*: MA > FA • *Attractiveness, competence*: FA > MA • *Persuasiveness (sport)*: MA > FA • *Persuasiveness (fashion)*: FA > MA
Zanbaka et al. (2006)	• 2 MA • 2 FA • 3-D • Voice • Realist • Speaker	• 41 MP • 97 FP • avg age 20.6	Listening to agents deliver a message to change participants' attitudes about university-wide comprehensive exams.	• Persuasiveness	• *Persuasiveness*: - MP: FA > MA - FP: MA > FA
Guadagno et al. (2007)	• 1 MA • 1 FA • 1 neutral • 3-D • Voice • Realist • Speaker	• 37 MP • 29 FP • avg age N/A	Listening to agents talk about changes to university security policy.	• Likeability • Credibility • Presentation quality • Persuasiveness	• *Likeability, credibility, presentation quality*: - MP: not significant - FP: FA > MA • *Persuasiveness*: - MP: MA > FA - FP: FA > MA
Guadagno et al. (2007)	• 1 MA • 1 FA • 3-D • Voice • Realist • Speaker	• 85 MP • 89 FP • avg age N/A	Listening to agents talk about changes to university security policy.	• Likeability • Credibility • Presentation quality • Social presence • Persuasiveness	• *Likeability*: FA > MA • *Credibility, presentation quality*: not significant • *Persuasiveness*: - MP: MA > FA - FP: not significant
Gulz et al. (2007)	• 2 MA • 2 FA • 2-D • Voice • Realist • Presenter	• 72 MP • 86 FP • avg age N/A	Listening to agents present university program engineering.	• Favorite agent • Interest	• *Favorite agent*: - MP: less feminine and less masculine agents > more masculine agent > more feminine agent - FP: less feminine and less masculine agents > more feminine agent > more masculine agent • *Interest*: - MP: more feminine and more masculine agents > less feminine and less masculine agents - FP: more femine and more masculine and less feminine agents > less masculine agent
Dill et al. (2008)	• 16 MA • 16 FA • 3-D • Realist • Video game characters	• 61 MP • 120 FP • avg age 18.82	Watching a PowerPoint presentation opposing still pictures of video game characters and male or female US senators. Reading a real-life story about the sexual harassment of a female student by a male professor.	• Tolerance for sexual harassment • Rape-supportive attitudes	• *Tolerance for sexual harassment*: MP > FP • *Rape-supportive attitudes*: MP > FP

(Continued)

TABLE 1 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Rosenberg-Kima et al. (2008)	• 1 MA • 1 FA • 3-D • Voice • Realist • Speaker	• 89 FP • avg age 19.7	Listening to an agent describe four female engineers and the benefits of engineering, with or without the agent present.	• Interest • Self-efficacy • Utility for engineering	• *Interest, self-efficacy, utility for engineering*: not significant
Rosenberg-Kima et al. (2008)	• 4 MA • 4 FA • 3-D • Voice • Realist • Speaker	• 111 FP • avg age 19.72	Listening to an agent describe four female engineers and the benefits of engineering, with or without the agent present.	• Interest • Self-efficacy • Utility for engineering • Fewer engineering gender stereotypes	• Self-efficacy and Interest in engineering: young and cool agents > other agents • *Utility for engineering*: MA > FA (not significant) • *Fewer engineering gender stereotypes*: FA > MA
Niculescu et al. (2009)	• 3 MA • 3 FA • 1 neutral • 3-D • Voice • Cartoon • Assistant	• 24 MP • 24 FP • avg age N/A	Interacting with agents about medical queries, evaluating an androgynous agent's gender either after or before rating non-androgynous agents.	• Androgynous agent's perceived gender	• *Androgynous agent's perceived gender*: - Non-androgynous agents rated first: - FP: more feminine - MP: more masculine - Androgynous agent rated first: - FP: more masculine - MP: more feminine
McDonnell et al. (2009)	• 1 MA • 1 FA • 2 neutral • 3-D • Realist • Subject	• 22 MP • 19 FP • avg age N/A	Watching a video of agents walking.	• Agents' perceived gender	• *Agents' perceived gender*: - FA (male walk): ambiguous - FA (neutral walk): female - MA (female walk): ambiguous - MA (neutral walk): male - Genderless agents: ambiguous - Genderless agents (female walk): female - Genderless agents (male walk): male - Genderless agents (neutral walk): female
McDonnell et al. (2009)	• 3 MA • 3 FA • 3-D • Realist • Subject	• 33 MP • 5 FP • avg age N/A	Watching a video of agents walking.	• Agents' perceived gender	• *Agents' perceived gender*: - FA rated 'most female': FA (bigger hips and breast size) > FA (smallest hips and breast size) - MA rated 'most male': no difference - Agents rated 'most ambiguous': FA (male walk) and MA (female walk)
Fox and Bailenson (2009)	• 4 FA • 3-D • Realist • Subject	• 43 MP • 40 FP • avg age 20.82	Participants encountered an agent (low gaze (LG) or high gaze (HG), masculine or feminine clothes) *via* virtual reality, then made judgments about them.	• Rape myth acceptance • Benevolent sexism • Hostile sexism	• *Rape myth acceptance*: masculine LG agent > feminine HG agent > masculine HG agent > feminine LG agent • *Benevolent sexism*: masculine LG agent > feminine LG agent > masculine HG agent

(Continued)

TABLE 1 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
					• *Benevolent sexism*: LG agent > HG agent • *Hostile sexism*: feminine HG agent > masculine HG agent
Cloud-Buckner et al. (2009)	• 2 MA • 2 FA • 3-D • Voice • Realist • Guide	• 19 MP • 16 FP • avg age N/A	Watching an agent introducing a college campus as an online tour guide.	• Friendliness • Anger • Cooperation • Self consciousness • Adventurousness • Sympathy • Sociability • Assertiveness • Cooperation • Self consciousness • Self discipline	• *Friendliness, anger, cooperation, self consciousness, adventurousness, sympathy*: Outgoing personality: MA > FA • *Sociability, assertiveness, cooperation, self consciousness, self discipline*: Introverted personality: FA > MA
Niculescu et al. (2010)	• 1 MA • 1 FA • 1 neutral • 3-D • Text • Cartoon • Tutor	• 4 MP • 4 FP • avg age N/A	Asking an agent medical questions.	• Comfortable • Confident • Less tense • Preferred agent	• *Comfortable, confident, less tense*: FA > MA and androgynous agent • *Preferred agent*: FA > MA > androgynous agent
Rosenberg-Kima et al. (2010)	• 2 MA • 2 FA • 3-D • Voice • Realist • Speaker	• 119 FP • avg age 21.49	Listening to an agent describe four female engineers and the benefits of engineering.	• Interest • Self-efficacy • Utility • Agent's likeability • Fewer engineering gender stereotypes	• *Interest*: - Black FP: Black FA > others - White FP: FA > MA • *Self-efficacy, utility, agent's likeability*: - Black FP (Black agents): FA > MA - Black FP (White agents): MA > FA • *Fewer engineering gender stereotypes*: - Black FP: Black agents > White agents - White FP: FA > MA
Astrid et al. (2010)	• 1 FA • 3-D • Voice • Realist • Questioner	• 41 MP • 42 FP • avg age 37.27	Answering personal questions from an agent.	• Weak • Shy • Naive • Compassionate • Inviting	• *Weak, shy, naive, compassionate, inviting*: not significant
Nunamaker et al. (2011)	• 1 MA • 1 FA • 3-D • Voice • Realist • Questioner	• 53 MP • 35 FP • avg age 25.45	Answering questions from an agent simulating an airport screening.	• Power • Trustworthiness • Expertise • Likability	• *Power, trustworthiness, expertise*: MA > FA • *Likability*: FA > MA
Kulms et al. (2011)	• 2 MA • 2 FA • 3-D • Voice • Realist • Questioner	• 32 MP • 40 FP • avg age 35.03	Answering casual questions asked by an agent, either in a low gaze (LG) or a high gaze (HG) condition.	• Masculinity • Positive evaluation • Social presence	• *Masculinity*: HG MA > LG MA • *Positive evaluation*: FA > MA • *Social presence*: MA > FA

(Continued)

TABLE 1 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Brahnam and De Angeli (2012)	• 8 MA • 8 FA • 3 neutral • 2-D • Text • Cartoon • Chatbot	• 127 MP • 73 FP • avg age N/A	Chatting over text with a chatbot.	• Sexual discourse • Avg number of words about money/job, and physical appearance	• *Sexual discourse, avg number of words (physical appearance)*: FA > MA • *Avg number of words (money/jobs)*: MA > FA (among adult agents)
Ozogul et al. (2013)	• 2 MA • 2 FA • 3-D • Cartoon • Tutor	• 35 MP • 42 FP • avg age 12.83	Rating pictures of agents.	• Gender preference • Preferred agent to learn about engineering from	• *Gender preference*: - FP: FA > MA - MP: MA > FA • *Preferred agent to learn about engineering from*: young FA > young MA > old MA > old FA
Payne et al. (2013)	• 4 MA • 4 FA • 2-D and 3-D • Cartoon • Assistant	• 220 MP • 358 FP • avg age 35.56	Choosing an agent to assist in self-service checkouts.	• Preferred agent	• *Preferred agent*: - FP: FA > MA - MP: MA > FA
Lunardo et al. (2016)	• 2 MA • 2 FA • 2-D • Text • Realist • Assistant	• 107 MP • 147 FP • avg age N/A	Interacting with an agent over text at fnac.com.	• Attractiveness	• *Attractiveness*: - Agents (corporate clothes): FA > MA - Agents (casual clothes): FA > MA (not significant)
van der Lubbe and Bosse (2017)	• 2 MA • 2 FA • 3-D • Voice • Realist • Employee	• 55 MP • 38 FP • avg age N/A	Interacting with an agent employee to negotiate the agent's salary (assertive agent or non-assertive agent).	• Appropriate language • Sensitive • No deal reached • Persuasiveness	• *Appropriate language*: assertive FA > assertive MA • *Sensitive*: non-assertive MA > non-assertive FA • *No deal reached*: assertive MA > assertive FA > non-assertive FA > non-assertive MA (not significant) • *Persuasiveness*: assertive FA > assertive MA
Feng et al. (2017)	• 1 MA • 1 FA • 3-D • Voice • Realist • Instructor	• 31 MP • 32 FP • avg age 21.37	Acting out a scene in presence of an agent giving negative feedback.	• Inspiration • Self-blame • Helpfulness • Preferred agent	• *Inspiration, self-blame, helpfulness, preferred agent*: FA > MA
Mell et al. (2017)	• 1 FA • 2-D • Text • Realist • Assistant	• 241 MP • 140 FP • avg age 35.13	Answering questions from a chatbot about sensitive information, either with a picture of a real woman, a picture of a female virtual agent, or no picture.	• Reported lies • Allowing the system to do a credit check • Providing their address	• *Reported lies*: human > no presence > agent • *Allowing the system to do a credit check*: - FP: agent > human > no presence - MP: no presence > agent > human • *Providing their address*: - FP: equal across conditions - MP: human > agent > no presence

(Continued)

TABLE 1 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Khashe et al. (2017)	• 1 MA • 1 FA • 3-D • Voice • Realist • Speaker	• 98 MP • 116 FP • avg age N/A	Requested to switch off the lights and open the window by a manager, either voice only, text only, or a virtual agent).	• Affectionate • Friendly • Likable • Persuasiveness	• *Affectionate, friendly, likable*: female (agent and voice only) > male (agent and voice only) • *Persuasiveness*: female (agent, voice only, text only) > male (agent, voice only, text only)
Kantharaju et al. (2018)	• 2 MA • 2 FA • 3-D • Voice • Realist • 2 experts • 2 motivators	• 113 MP • 92 FP • avg age N/A	Listening to a persuasive conversation about cinema between agents.	• Distant • Arrogant • Forceful • Credible • Persuasiveness	• *Distant, arrogant, forceful, credible, persuasiveness*: MP > FP
Akbar et al. (2018)	• 1 MA • 1 FA • 3-D • Text • Realist • Interviewer	• 158 MP • 158 FP • avg age N/A	Interviewed by an agent over text for a job in a financial firm.	• Agreeableness • Trustworthiness	• *Agreeableness*: opposite gender agent > matching-gender agent • *Trustworthiness*: matching-gender agent > opposite gender agent
Mousas et al. (2018)	• 2 MA • 3-D • Realist • Subject	• 56 MP • 16 FP • avg age 23.24	Answering questions about the agents (e.g., "Would you feel uneasy if this virtual character communicated with you?") by the experimenter while the agent walked toward the participant.	• Easiness • Comfortableness • Readiness for interaction • Likeability	• *Easiness, comfortableness, readiness for interaction*: MP > FP • *Likeability*: - zombie agent: MP > FP - MA: not significant
Ait Challal and Grynszpan (2018)	• 1 MA • 1 FA • 3-D • Realist • Subject	• 12 MP • 12 FP • avg age 23.6	Watched virtual agents sit in front of them (in gaze following, gaze avoidance, high direct gaze, and low direct gaze conditions). Judging their personalities.	• Neuroticism • Agreeableness	• *Neuroticism*: FA > MA • *Agreeableness (high direct gaze condition)*: MA > FA
ter Stal et al. (2020)	• 4 MA • 4 FA • 2-D • Cartoon • 4 experts, 4 peers	• 67 MP • 69 FP • avg age 51.36	Observing and rating 8 agents.	• Friendliness • Expertise • Authority	• *Friendliness*: FA > MA • *Expertise, authority*: MA > FA
ter Stal et al. (2020)	• 4 MA • 4 FA • 2-D • Cartoon • 4 experts, 4 peers	• 35 MP • 30 FP • avg age 67.85	Observing and rating 8 agents.	• Friendliness • Authority	• *Friendliness*: not significant • *Authority*: MA > FA
Zibrek et al. (2020)	• 2 neutral • 3-D • Realist • Subject	• 10 MP • 10 FP • avg age N/A	Pressing a button as soon as they felt uncomfortable with the distance between themselves and an agent walking toward them.	• Genderless agents' perceived gender	• *Genderless agents' perceived gender*: - female motions = female - male motions = male

(Continued)

TABLE 1 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Richards et al. (2020)	• 6 MA • 6 FA • 3-D • Voice • Realist • Assistant	• 43 MP • 146 FP • avg age 21.7	Watching 12 videos of 12 different agents introducing themselves.	• Favorite agent (before and after watching the videos)	• *Favorite agent (before)*: - FP: gender does not matter > FA > MA - MP: gender does not matter > MA > FA • *Favorite agent (after)*: Mediterranean FA > Asian FA > White FA
Nag and Yalçın (2020)	• 1 MA • 1 FA • 1 neutral • 3-D • Realist • Subject	• 41 MP • 31 FP • avg age 21.7	Looking at pictures of agents and rating them.	• Communion • Agency • Competence	• *Communion*: FA > MA (not significant) • *Agency, competence*: not significant
Esposito et al. (2021)	• 2 MA • 2 FA • 3-D • Voice • Realist • Assistant	• 22 MP • 24 FP • avg age 71.59	Watching a video of an agent talking about daycare facilities for the elderly.	• Willingness to interact with the agent • Attractiveness • Usefulness • Presentable • Professional • Of good taste • Pleasant • Original • Creative • Captivating	• *Willingness to interact with the agent, attractiveness, usefulness, presentable, professional, of good taste, pleasant, original, creative, captivating*: FA > MA
Esposito et al. (2021)	• 2 MA • 2 FA • 3-D • Voice • Realist • Assistant	• 20 MP • 25 FP • avg age 71.22	Watching a video of an agent talking about daycare facilities for the elderly. (2nd experiment).	• Willingness to interact with the agent • Attractiveness • Usefulness • Presentable • Professional • Of good taste • Pleasant • Original • Creative • Captivating	• *Presentable, professional, of good taste, pleasant*: FA > MA • *Willingness to interact with the agent, attractiveness, usefulness, original, creative, captivating*: Not significant
Vilaro et al. (2021)	• 3 FA • 3-D • Voice, text • Realist • Assistant, expert	• 53 FP • avg age 60.90	Watching an agent deliver colorectal cancer screening messages.	• Trustworthiness • Expertise	• *Trustworthiness*: not significant • *Expertise*: agents (white medical coat) > agent (casual clothes)
Antonio Gómez-Jáuregui et al. (2021)	• 1 MA • 1 FA • 3-D • Voice • Realist • Interviewer	• 16 MP • 16 FP • avg age 29.95	Introducing themselves to a blurred-face virtual agent for a job interview.	• Dominance • Warmth	• *Dominance*: not significant • *Warmth*: FA (mirrored movements) > FA (random movements)

(Continued)

TABLE 1 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Świdrak et al. (2021)	• 1 MA • 1 FA • 3-D • Voice • Realist • Player	• 15 MP • 19 FP • avg age 25	Playing a negotiation/ decision-making game with a female and a male agent.	• Touch pleasantness • Touch awkwardness • Touch adequacy • Persuasiveness	• *Touch pleasantness*: FA > MA • *Touch awkwardness*: FP > MP • *Touch adequacy*: FA perceived as more masculine > FA perceived as less masculine • *Persuasiveness*: - MP: agents perceived as more masculine > agents perceived as less masculine - FP: depends on the offer
Świdrak et al. (2021)	• 2 MA • 2 FA • 3-D • Voice • Realist • Player	• 40 MP • avg age 23	Playing a negotiation/ decision-making game with two female and two male agents.	• Masculinity • Touch pleasantness • Touch awkwardness • Touch adequacy • Persuasiveness	• *Masculinity*: masculine FA > feminine MA • *Touch pleasantness*: FA > feminine MA • *Touch awkwardness*: feminine MA > others (not significant) • *Touch adequacy*: others > feminine MA (not significant) • *Persuasiveness*: masculine-perceived agents > feminine-perceived agents

Articles are listed from oldest to most recent. FA, female agent; MA, male agent; FP, female participants; MP, male participants. The agents' column describes the number of agents depending on their gender, their dimension (2-D or 3-D), their appearance (realist or cartoon), and their role. The participants' column describes the number of men and women who participated in the study and the average age. In the result(s) column, "MP > FP" means it impacted more the male participants than the female participants. "FA > MA" means the female agent has more impact than the male agent. Explanations are in Section 3.4.

tables in Section 3.4 and Section 3.5. In Section 3.4, we address research highlighting the impact of virtual agents' gender on users' perceptions. In Section 3.5, we focus on pedagogical agents and the impact of their gender on learners' academic outcomes. The last section discusses what could be done in future research on virtual learning environments to reduce gender stereotypes and improve learners' performance, and the important research questions that arise from this review.

2. METHODS

2.1. Search Strategy

This article examines research on the impact of virtual agents' gender on learners but also more generally on users' behavior and perceptions. For this purpose, we reviewed articles from the Web of Science database over 21 years from 2000 to 2021. To collect the relevant studies, we conducted an online database search with the query *gender+("virtual agent*" OR "virtual character*")*. This systematic review was conducted according to the PRISMA guidelines presented in **Figure 1** (Webster and Watson, 2002) as follows: (1) scanning databases and starting with the major contributions in the leading journals, (2) reviewing the citations for the articles identified in step 1 to determine prior articles that should be considered, and (3) identifying articles citing the key articles identified in the previous steps. We used Google Scholar for the last step. A total of 120 articles were retained after following these steps.

2.2. Selection of Articles

From this set of articles, we selected empirical studies analyzing the effect of virtual agents' gender on users' perceptions, behaviors, and academic outcomes. We only took into account embodied virtual agents (i.e., we excluded studies on vocal assistants). We focused on Western culture and, thus, only selected papers relating to this culture. We eliminated articles only about avatars (users embodying a virtual agent) which were mainly about video games. In the end, we retained a set of 59 articles. We distinguished two types of research articles: Perceptive studies of virtual agents depending on their gender, regardless of the application domain, and research studies on the impact of gendered virtual agents in the context of a learning task.

This systematic review focuses on how virtual agents are designed, and the impact of their gender on different academic outcomes (motivation, learning, interest), but also on participants' perceptions of the agents. The research questions guiding this review are as follows:

1. Do students perceive virtual agents differently depending on their own gender and the gender of the agent?
2. Does the gender of pedagogical agents influence students' academic performance and self-evaluations?

TABLE 2 | Summary of research studies on the impact of gendered virtual agents in the context of a learning task.

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Moreno et al. (2002)	• 2 MA • 2 FA • 3-D • Voice • Realist • Tutor	• 12 MP • 27 FP • avg age 20	Watching a video of a virtual agent giving a course, taking a multiple-choice test.	• Performance • Perceived masculinity, femininity	• *Performance*: MA > FA • *Perceived masculinity, femininity*: - FA: very feminine - MA: masculine
Baylor and Kim (2004)	• 4 MA • 4 FA • 2-D, 3-D • Voice • Realist, cartoon • Tutor	• 94 MP • 218 FP • avg age 20.54	Creating an instructional schedule with a virtual agent's help.	• Self-efficacy • Self-regulation • Knowledgeability • Intelligence • Learning	• *Self-efficacy, self-regulation, knowledgeability, intelligence*: MA > FA • *Learning*: not significant
Baylor and Kim (2004)	• 6 MA • 6 FA • 2-D, 3-D • Voice • Realist, cartoon • Expert, motivator, mentor	• 89 MP • 140 FP • avg age 19.39	Creating an instructional planning with a virtual agent's help.	• Knowledgeability • Intelligence • Learning • Self-regulation • Self-efficacy	• *Knowledgeability, intelligence*: MA > FA • *Learning, self-regulation:* not significant • *Self-efficacy*: FA > MA
Moreno and Flowerday (2006)	• 5 MA • 5 FA • 2-D • Voice • Realist • Tutor	• 21 MP • 59 FP • avg age 26.88	Watching a video of a course taught by a virtual agent, taking a test.	• Helpfulness • Motivation • Selected agent • Learning	• *Helpfulness, motivation, learning*: not significant • *Selected agent*: matching-gender agent = opposite gender agent
Kim et al. (2007)	• 1 MA • 1 FA • 3-D • Voice • Realist • Companion	• 11 MP • 45 FP • avg age 20.71	Creating a course on economic concepts with a virtual agent's help.	• Facilitating learning • Engaging • Human-like • Learning (recall)	• *Facilitating learning, engaging, human-like*: MA > FA • *Learning (recall)*: not significant
Plant et al. (2009)	• 1 MA • 1 FA • 3-D • Voice • Realist • Speaker	• 45 MP • 61 FP • avg age 13.63	Listening to a story about four female engineers and the benefits of engineering, either delivered by an agent or voice-only. Taking a math test.	• Interest • Utility • Self-efficacy • Performance • Fewer engineering gender stereotypes	• *Interest, utility*: FA > MA and no agent • *Self-efficacy*: MA and FA > no agent • *Performance*: FA > MA • *Fewer engineering gender stereotypes*: - MP: agents > no agent - FP: FA and no agent > MA
Hayes et al. (2010)	• 1 MA • 1 FA • 3-D • Voice • Realist • Observer	• 35 MP • avg age 19.77	Controlling an avatar (1st or 3rd person view) while taking a math test, in the presence of a male or female agent, or without an agent.	• Social presence • Performance • Response times	• *Social presence*: MA > FA and no agent • *Performance, response times*: - 1st person: no agent and MA > FA - 3rd person: FA > no agent and MA

(Continued)

TABLE 2 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Kim and Wei (2011)	• 2 MA • 2 FA • 3-D • Voice • Realist • Tutor	• 110 MP • 100 FP • avg age 15.93	Taking a math test without an agent, watching an agent explaining the lessons, resolving math problems with the agent (training), taking a 2nd math test without an agent.	• Selected agent • Performance	• *Selected agent*: matching gender and matching ethnicity agents > others • *Performance*: everyone improved
Silvervarg et al. (2013)	• 1 MA • 1 FA • 1 neutral • 2-D • Text • Cartoon • Tutee	• 46 MP • 37 FP • 12–14 years old	Interacting with an androgynous virtual tutee on a math lesson, then with either a female or a male virtual tutee.	• Perceived androgyny • Preferred agent as tutee • Preferred agent as chat partner	• *Perceived androgyny*: androgynous agent = androgynous • *Preferred agent as tutee*: - FP: androgynous agent > MA and FA - MP: androgynous agent > MA and FA (not significant) • *Preferred agent as chat partner*: - FP: androgynous agent > MA - MP: androgynous agent > FA
Kim and Lim (2013)	• 2 FA • 3-D • Voice • Realist • Tutor	• 64 MP • 56 FP • avg age 15.93	Taking a math test without an agent, learning lessons with or without an agent, resolving math problems with or without an agent (training), taking a 2nd math test without an agent.	• Performance • Self-efficacy	• *Performance*: everyone improved • *Self-efficacy*: - FP: agent present > no agent - MP: no increase
Kim (2013)	• 1 MA • 1 FA • 1 neutral • 2-D, 3-D • Voice, text • Cartoon • Tutor	• 68 MP • 73 FP • avg age N/A	Answering questions about a text asked by a virtual agent.	• Text comprehension	• *Text comprehension*: - FP = MP - FP: MA and FA > robot agent - MP: MA > FA and robot agent
Johnson et al. (2013)	• 1 MA • 1 FA • 3-D • Voice • Cartoon • Tutor	• 88 MP • 109 FP • avg age 12.1	Watching an agent teaching a lesson on electrical circuits, taking a multiple-choice test.	• Performance • Program evaluation	• *Performance*: not significant • *Program evaluation*: FA > MA
Ozogul et al. (2013)	• 2 MA • 2 FA • 3-D • Voice • Cartoon • Tutor	• 173 MP • 161 FP • avg age 12.3	Watching an agent (chosen or randomly assigned) teaching a lesson on electrical circuits, taking a multiple-choice test.	• Performance • Program evaluation • Selected agent	• *Performance*: - Random agent: not significant - Selected agent: opposite gender agent > matching-gender agent • *Program evaluation*: not significant • *Selected agent*: matching-gender agent > opposite gender agent
Shiban et al. (2015)	• 1 MA • 1 FA • 3-D • Text • Cartoon • Tutor	• 21 MP • 73 FP • avg age 20.20	Taking a math test while a virtual agent provided feedback.	• Interest • Motivation • Enjoyment • Credible • Engaging • Human-like • Facilitating learning • Performance	• *Interest, motivation*: FA > MA • *Enjoyment*: not significant • *Credible, engaging, human-like*: MA > FA • *Facilitating learning*: not significant • *Performance*: MA > FA (slightly)

(Continued)

TABLE 2 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Kim (2016)	• 2 MA • 2 FA • 3-D • Voice • Realist • Tutor	• 67 FP • avg age 15.51	Listening to an agent speak persuasively about the benefits of STEM fields, solving math problems with the same agent, solving math problems without the agent.	• Credibility • Friendliness • Helpfulness • Positive attitudes to learn math	• *Credibility, friendliness, helpfulness*: - Ethnic-minority participants: - MA: peer agent > teacher agent - FA: teacher agent > peer agent - Caucasians participants: - MA and FA: not significant • *Positive attitudes to learn math*: - Ethnic-minority participants: - MA: peer agent > teacher agent - FA: teacher agent > peer agent - Caucasians participants: - MA and FA: not significant
Krämer et al. (2016)	• 2 MA • 2 FA • 3-D • Voice • Realist • Motivating interviewer	• 60 MP • 68 FP • avg age 23.85	Taking a math test without an agent, then taking a math test with an agent present explaining the procedure.	• Motivation • Sense of rapport • Performance	• *Motivation, sense of rapport*: not significant • *Performance*: - FP and rapport agent: MA > FA - MP and rapport agent: FA > MA
Li et al. (2016)	• 1 MA • 1 neutral • 3-D • Voice • Realist • Tutor	• 20 MP • 20 FP • avg age 20.48	Watching an agent present slides on courses about Human-Computer Interaction.	• Learning	• *Learning*: - MP: agent robot > real human (male) > MA > still image of a robot - FP: no differences
Jeong et al. (2017)	• 1 MA • 1 FA • 3-D • Voice • Realist • Instructor	• 54 MP • 63 FP • avg age 20.94	Listening to negative feedback from an instructor agent while acting out a scene. Reproducing the scene with the instructor agent and a student agent (no feedback).	• Moving forward • Moving backward	• *Moving forward*: - FP: FA > MA - MP: MA > FA • *Moving backward*: - FP: MA > FA - MP: FA > MA
Pezzullo et al. (2017)	• 1 FA • 3-D • Voice • Realist • Companion	• 54 MP • 63 FP • avg age 13.30	Playing a game about biology courses with a virtual agent's help.	• Mental demand • Engagement with the agent • Performance	• *Mental demand, engagement with the agent*: FP > MP • *Performance*: FP = MP
Wirzberger et al. (2019)	• 1 MA • 3-D • Voice • Realist • Instructor	• 27 MP • 35 FP • avg age 69.03	Memorizing a word list after taking a memory training course led by an agent.	• Learning (recall)	• *Learning (recall)*: FP > MP
Makransky et al. (2019)	• 1 FA • 1 neutral • 3-D • Voice • Realist • Tutor	• 33 MP • 33 FP • avg age N/A	Watching a virtual agent teaching laboratory safety, taking tests.	• Social presence • Learning (recall and transfer-learning)	• *Social presence*: - FP: FA = drone agent - MP: FA > drone agent • Learning (recall and *transfer-learning)*: - FP = MP - FP: FA > drone agent - MP: drone agent > FA

(Continued)

TABLE 2 | Continued

Reference(s)	Agent(s)	Participant(s)	Task(s)	Measure(s)	Result(s)
Chang et al. (2019)	• 1 MA • 3-D • Voice • Realist • Instructor	• 76 FP • avg age N/A	Controlling either a male or a female avatar, learning how to solve arithmetic problems from a male agent (either a dominant or a non-dominant agent, based on his body posture), solving problems without the agent present.	• Learning (recall and performance)	• *Learning (recall and performance)*: - non-dominant agent > dominant agent - No significant effect of avatar's gender
Sajjadi et al. (2020)	• 1 MA • 1 FA • 3-D • Voice • Realist • Instructor	• 8 MP • 4 FP • avg age 19.6	Observing geologic formations in a virtual environment, answering questions asked by an agent.	• Perceived learning effectiveness • Learning • Leadership • Friendliness • Social and spacial presence	• *Perceived learning effectiveness*: FA > MA • *Learning*: not significant • *Leadership, friendliness, social and spacial presence*: FA > MA (not significant)
Spilioto-poulos et al. (2020)	• 1 FA • 3-D • Voice • Realist • Tutor	• 24 MP • 16 FP • avg age 20	Learning how to use argumentation, how to be empathetic to the needs of others, how to reach agreements through negotiation with a virtual agent.	• Self-efficacy • System easiness • Helpfulness • Learning	• *Self-efficacy*: not significant • *System easiness*: FP > MP • *Helpfullness*: MP > FP • *Learning*: not significant (increase overall)

Articles are listed from oldest to most recent. FA, female agent; MA, male agent; FP, female participants; MP, male participants. The agents' column describes the number of agents depending on their gender, their dimension (2-D or 3-D), their appearance (realist or cartoon), and their role. The participants' column describes the number of men and women who participated in the study and the average age. In the result(s) column, "MP > FP" means it impacted more the male participants than the female participants. "FA > MA" means the female agent has more impact than the male agent. Explanations are in Section 3.5.

3. Are there tasks or academic situations to which a male virtual agent is better suited than a female virtual agent, and vice versa, according to empirical evidence?
4. How do a virtual agent's pedagogical roles impact these results?
5. How do a virtual agent's appearance and interactive capacities impact these results?
6. Are androgynous virtual agents a potential solution to combatting gender stereotypes?

3. SYSTEMATIC REVIEW OF VIRTUAL AGENTS' GENDER AND ITS IMPACT ON USERS' PERCEPTIONS AND ACADEMIC OUTCOMES

In this section, we first review the different measures used to assess users' perceptions of agents, users' learning, and self-evaluations. We then highlight the persistence of gender stereotypes in human-machine interactions by presenting research on users' perceptions of virtual agents depending on their gender (Section 3.4). Second, we focus on pedagogical agents and discuss research that shows the effect of their gender on learners (Section 3.5).

3.1. Subjective Measures of Users' Perceptions of Agents'

Most of the studies used post-experience questionnaires to assess users' perceptions of virtual agents. Likert scale items were used to determine participants' stereotyped attributions of the agents, corresponding to communal traits stereotypically associated with women (e.g., affectionate, compassionate, sensitive, inviting, helpful), agency traits stereotypically associated with men (e.g., arrogant, ambitious, aggressive, courageous, and decisive), and competence traits associated more often with men (e.g., knowledgeable, intelligent, expert, credible, creative, innovative and organized) (Lee, 2003; Nunamaker et al., 2011; Feng et al., 2017; Khashe et al., 2017; van der Lubbe and Bosse, 2017; Kantharaju et al., 2018; Sczesny et al., 2018). Another questionnaire was sometimes used to determine which stereotypical gendered traits users applied to the agents (Kulms et al., 2011). This scale, the Bem Sex Role Inventory (BSRI) developed by Bem (1974), measures the construction of the gender schema of individuals, aims to highlight androgyny, and questions the usual dichotomy of female/male gendered traits stereotypically attributed to people. The BSRI consists of 20 positive items stereotypically associated with men (e.g., independent, analytical), 20 other positive items stereotypically associated with women (e.g., compassionate, loves children), and 20 other positive neutral items (e.g., tactful, reliable). The agents' gender perception was evaluated with a 5-point Likert sliding

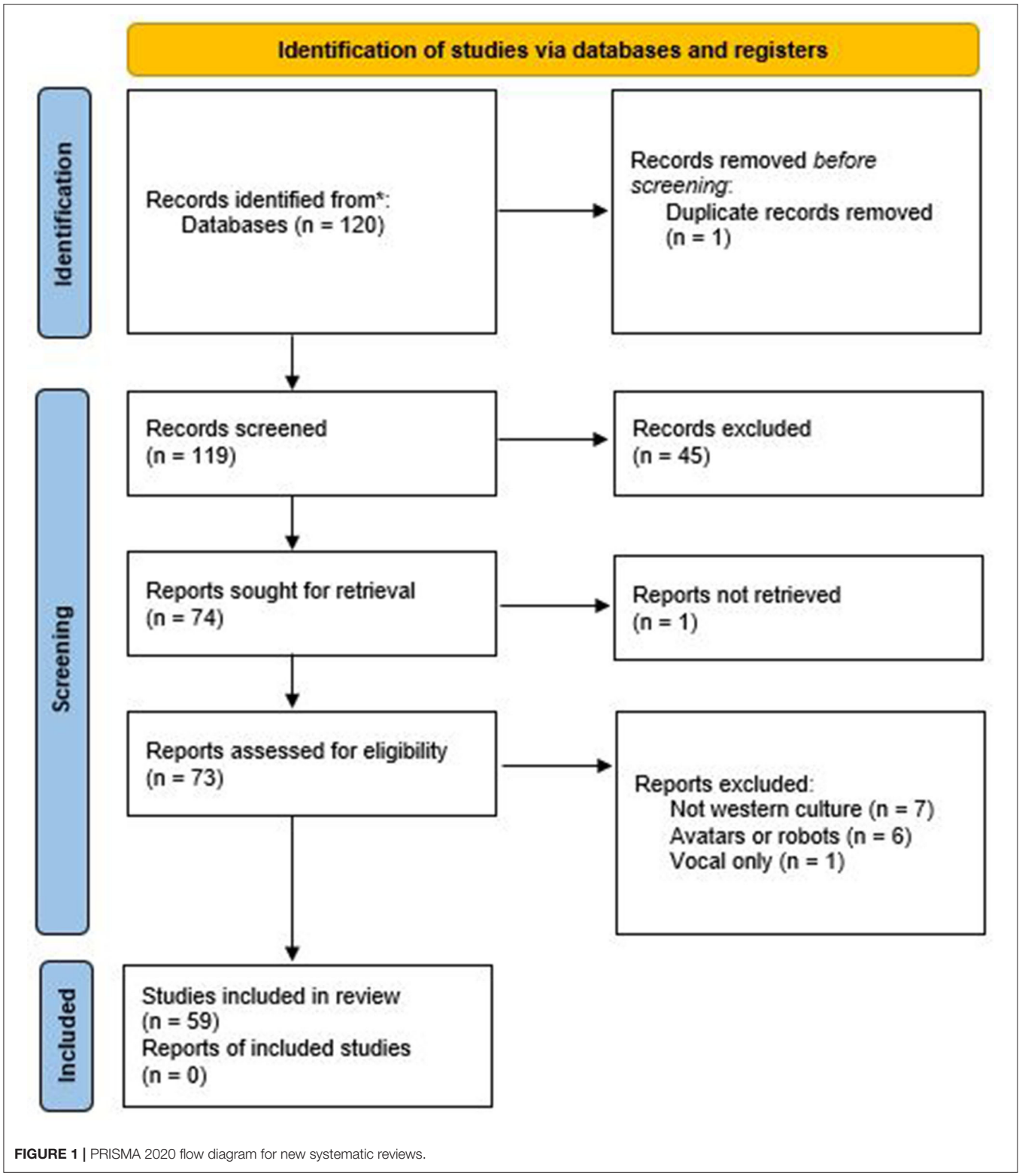

FIGURE 1 | PRISMA 2020 flow diagram for new systematic reviews.

scale, e.g., with men=1, androgynous=4, and women=7 (Lee, 2003; McDonnell et al., 2009; Niculescu et al., 2009; Nag and Yalçın, 2020). Other social attitude perceptions were also assessed with Likert scale items, such as the perceived friendliness, trustworthiness, likability, and social presence of the agent (Lee, 2003; Guadagno et al., 2007; Nunamaker et al., 2011; Lunardo et al., 2016; Khashe et al., 2017; Akbar et al., 2018). Social presence is particularly important as it provides individuals with

the possibility of developing a relationship or having a social interaction with one another, as they recognize each other as "social beings" (Biocca et al., 2003). Social presence is commonly defined as the sensation of being in the presence of a real person and having access to their feelings (Biocca, 1997), and can be assessed with a 5-item survey (e.g., "I feel that the person is watching me and is aware of my presence") (Bailenson et al., 2001) and the Networked Minds Questionnaire (e.g., "The other individual didn't notice me in the room") (Biocca et al., 2001), as used by Kulms et al. (2011).

3.2. Objective Measures of Learning

The impact of pedagogical agents on users' learning can be assessed by measuring users' performance in an exercise by comparing different conditions: for example, virtual agents with different behaviors (Chang et al., 2019), the presence of gendered virtual agents (Kim and Wei, 2011), or virtual agents with different genders (Kim, 2013). Performance can be measured with different problem-solving tests: using knowledge retention (using past knowledge to solve a problem, Sajjadi et al., 2020), recall (the ability to remember items, Wirzberger et al., 2019), or transfer learning (using past knowledge to solve new problems, Makransky et al., 2019). In addition to performance, researchers can also evaluate response times and effort. Effort can be measured by comparing the number of problems solved (that are not necessarily correct) in different problem-solving tests (Krämer et al., 2016). Response times correspond to the duration required to solve a problem (Hayes et al., 2010).

3.3. Users' Self-Evaluations

In learning situations, other more subjective measures than performance are rated using Likert-scale items. These measures include the interest in a task or a domain, (e.g., "I will take a hard sciences course as an elective," Rosenberg-Kima et al., 2008), beliefs about the utility of a task or a domain (e.g., "I would have many good career opportunities if I was a hard science major," Plant et al., 2009), learners' self-efficacy as in feeling capable of performing a task (e.g., "I can achieve high grades in math," Kim and Wei, 2011), learners' self-regulation to regulate their behaviors to succeed in a task (e.g., "I kept track of my progress," Baylor and Kim, 2004), learners' motivation assessed with the Situational Motivation Scale (SIMS) (Guay et al., 2000) which includes 16 items about the motivation to work on tasks (e.g., "Because I am doing it for my own good," Krämer et al., 2016), learners' enjoyment (e.g., "How much did you enjoy preparing for the exam?," Shiban et al., 2015), their perceived learning effectiveness (e.g., "I gained a good understanding of the basic concepts of the materials," Sajjadi et al., 2020), and their mental demand to know how much mental and perceptual activity was required (thinking, deciding, calculating, etc.), e.g., "Was the task easy or demanding?" (Hart and Staveland, 1988; Pezzullo et al., 2017).

3.4. Evidence of the Persistence of Gender Stereotypes in Human-Machine Interactions

We have summarized the selected studies on users' perceptions of virtual agents depending on their gender in each line of the following table. We stated agents' characteristics, the number of male and female participants with their average age, tasks of the study, the observed measures, and the study's results. Some acronyms are present in this table. We used MA for Male Agent(s) and FA for Female Agent(s). In the same logic, MP is used for Male Participant(s), and FP for Female Participant(s).

The studies presented in **Table 1** show that gender stereotypes persist in human-machine interactions. Users' behavior varies according to the gendered appearance of virtual agents. For example, in De Angeli and Brahnam (2006), the female virtual agent received several violent sexual propositions and even rape threats; the male virtual agent received only a few sexual propositions, none of them violent ("gently presses my lips to yours into a small kiss"), and the other sexual comments made during the interactions with the male virtual agent targeted his girlfriend. In a similar study by Brahnam and De Angeli (2012), users interacted with several pairs of female/male agents, including child agents, White agents, Black agents, and "old" agents. The female agents were the target of significantly more sexual discourse, comments on their appearance, and swear words than the male agents; this was even true for the pair of child agents. Other features of agents influenced the conversational topics, such as their age and appearance: users talked more about jobs, achievements, and money with old agents dressed in formal clothing than with any other pair of agents. However, gender stereotypes still applied to this category, since users interacted more with the older male agent about these topics than with the older female agent.

An agent's gender also has a direct influence on participants' decisions. For instance, Lee (2003) reported that users followed more advice from virtual agents when their gender stereotypically matched the topic (e.g., a female agent and cosmetics, a male agent and sports). In this study, the female virtual agent presented as particularly feminine. This result should, thus, be verified in a separate study using a female virtual agent presenting a sport-oriented appearance to determine whether these results are due solely to gender and not to the agents' presentation (clothes and make-up). In Guadagno et al. (2007), the male virtual agent was more persuasive when perceived to be computer-controlled rather than human-controlled. The opposite was true for the female virtual agent. The authors concluded that these results may have been due to gender stereotypes, specifically by the "participants" expectations for interacting with a computer being more consistent with masculine stereotypes (e.g., competent), whereas expectations for interacting with a human are more consistent with feminine stereotypes (e.g., warm)."Not only the gender of a virtual agent but even their perceived masculinity can influence participants' decisions. In a decision-making game where virtual agents made a monetary offer to male participants, the number of offers accepted was higher with the agents perceived as more masculine (Świdrak et al., 2021). The same

results were obtained in a similar study for male participants; in contrast, female participants accepted more offers from the agents than male participants but were only influenced by the offer itself, not by the agents' perceived masculinity (Świdrak et al., 2021).

An agent's gender also has an impact on how users perceive the agent in terms of stereotypical traits attributed to men and women (Sczesny et al., 2018). In a study by Nunamaker et al. (2011), the male agent was perceived as more powerful, whereas the female agent was perceived as more likable. Even when male and female agents wore the same clothes, exhibited the same verbal and non-verbal behaviors, and had their faces blurred (thus lacking salient indicators of gender), the female agent was rated higher for warmth than the male agent; however, they were rated similarly for dominance, a trait typically associated with men (Antonio Gómez-Jáuregui et al., 2021). In contrast, Kulms et al. (2011), found in their main experiment that participants did not ascribe more masculine traits to the male agents nor more feminine traits to the female agents, unlike in their pretest with 14 participants using still pictures of the same virtual agents. The authors concluded that stereotyped attributions became less important when participants could interpret the behavior of the agents. However, a study by Ait Challal and Grynszpan (2018) contradicts this conclusion: the female agent was rated as less agreeable than the male agent when using high direct gaze. The authors suggested that participants were less tolerant of dominance when expressed by a female agent. Gender stereotypes associated with users' gender can also impact the ratings of virtual agents. In a study by Mousas et al. (2018), male participants reported feeling more at ease and comfortable with a zombie agent than female participants; they also liked the zombie agent more than the female participants did. The authors concluded that gender stereotypes may have influenced the results because stereotypes call for men to be calmer in the face of fear and embarrassment/disgust and to report milder emotional reactions.

Contexts stereotypically associated with one gender may also have an impact on participants' preferences as to the gender of agents: in two experiments conducted by ter Stal et al. (2020), elderly participants preferred still pictures of female agents in a healthcare context. According to the authors, this result could be due to the task—health coaching—being associated with female gender stereotypes. In addition, male agents were rated as more authoritarian and expert than female agents. In a study by Gulz et al. (2007), when virtual agents presented university programs in computer engineering, participants' interest was higher in feminine and masculine agents as compared to "neutral" agents (a less feminine female agent and a less masculine male agent). However, participants who ranked the less feminine female agent as the best presenter chose her because they believed that she could make more girls interested in computer engineering ("she seems young and nice, and I think she would make more girls interested"); and participants who ranked the feminine female agent as the worst presenter chose her because she was a woman who did not seem to belong in that context ("as I said, a woman feels more welcoming than a man, but she looked so styled, which I don't like"). These results show that gender stereotypes apply to the appearance of female agents.

In addition to context, agents' roles can also influence how users perceive them. When female agents were presented as assistants to elderly people in their daily life, participants found them to be more worth interacting with, more useful, efficient, and well designed, and more captivating, exciting, engaging, and attractive than male agents (Esposito et al., 2021). However, in a similar experiment with silent agents, the agents' gender did not affect the participants in terms of the same criteria (Esposito et al., 2021). Voices could have influenced the perceived agents' masculinity/femininity, but this was not measured in the studies. In a different study, expert agents were rated as more credible than motivational agents regardless of their gender (Kantharaju et al., 2018).

However, a recent study by Nag and Yalçın (2020) contradicts previous research on how humans perceive virtual agents depending on their gender: still pictures of male and female agents were generally rated similarly for *agency* (traits typically associated with men: *ambitious, aggressive, courageous, decisive*) and *competence* (traits typically associated with men: *creative, intelligent, innovative, organized*), but not for *communion* (traits typically associated with women: *affectionate, compassionate, sensitive, inviting, helpful*) where female agents were rated higher. A limitation of this study is that the female and male agents were quite similar in appearance. This being said, the results of the study tend to be coherent with the evolution of gender stereotypes reported by Eagly et al. (2020) for the perception of *agency* and *competence* traits perception in interpersonal interactions: the gap in *agency* and *competence* in favor of men has reduced. However, the *communion* traits are still largely attributed to women. This raises the question of whether the evolution in the perception of stereotypes in human-human interactions shown by Eagly et al. (2020) can be observed similarly in human-virtual agent interactions.

Based on the research presented above, it seems that male virtual agents are perceived as more competent, especially regarding stereotypically male-related topics. They appear as better suited to represent a pedagogical virtual tutor in STEM fields (Science, Technology, Engineering, and Mathematics) since these fields are perceived as masculine (Makarova et al., 2019). In the next section, we focus more specifically on research on pedagogical agents and the impact of their gender on users' academic outcomes.

3.5. The Effect of Virtual Agents' Gender on Academic Outcomes

We have summarized the selected studies on the impact of gendered virtual agents in the context of a learning task in each line of the following table. We stated agents' characteristics, the number of male and female participants with their average age, tasks of the study, the observed measures, and the study's results. Some acronyms are present in this table. We used MA for Male Agent(s) and FA for Female Agent(s). In the same logic, MP is used for Male Participant(s), and FP for Female Participant(s).

Various studies on virtual learning environments (**Table 2**) have reported that the gender of a pedagogical agent may have an impact on the learning performance of users. In a recent article, Makransky et al. (2019) showed that young girls performed better on scientific tasks (in terms of learning and transfer learning) when taught by a virtual female scientist than by a virtual drone. The opposite was true for boys. The researchers argued that boys identified with the drone, while girls identified with the female agent. However, research opposing human-like vs. robot-like agents does not take into account other factors that may influence how girls learn. In a study by Shiban et al. (2015), female learners were more motivated and interested in math when trained by a female agent as compared to a male agent. However, they obtained better results with the male agent, which may be explained by their perception of the agents' appearances: the male agent was older and wore a tie, while the female agent was young and pretty. According to the authors, the participants' performance improved because the male agent was perceived as an expert, and virtual agents perceived as experts have been shown to improve learners' performance (Baylor and Kim, 2004). The researchers also concluded that the female participants' motivation and interest improved with the female agent because there were more female participants in the study and because of the agent's similarity (in age and gender) to them, in line with the "similarity hypothesis" (also found in Rosenberg-Kima et al. 2008). This argument is supported by Bandura's social cognitive learning theory: people often learn by imitating people whom they perceive as similar (or superior: higher in rank or status) to them and who are, therefore, accepted as social role models (Bandura and National Inst of Mental Health, 1986). This theory also bears out in a study by Plant et al. (2009), where a female agent raised participants' self-efficacy by delivering a message on the benefits of engineering, resulting in better performance and more interest in math.

However, other research has demonstrated a positive effect of male agents as compared to female agents in pedagogical tasks. For instance, in two experiments conducted by Baylor and Kim (2004), a virtual agent helped participants create a schedule. The agent's gender did not impact learning but did affect self-efficacy, which increased more in the first experiment with the male agent than the female agent; the contrary occurred in the second experiment for both male and female participants. The researchers suggested that there was a bias in the first experiment, as participants rated the male agent as more interesting and useful than the female agent. In the second experiment, participants viewed the female agent as less expert and knowledgeable than the male agent, despite receiving the same instructions from both agents; some research has indicated that agents perceived as less intelligent could lead to greater self-efficacy (Baylor and Kim, 2005). In a similar experiment by Kim et al. (2007), the researchers introduced a female and a male pedagogical agent to help students design an e-learning course, which included creating a schedule. Students working with the male agent rated him higher on facilitating learning, being engaging, and being human-like than students working with the female agent. Notably, the male agent had a more positive impact than the female agent on the participants' interest and learning in terms of recall (the ability to remember what the agent said during the task).

Other factors may also come into play in studies on the impact of virtual agents on learning. In Moreno et al. (2002), participants watched a video of a virtual agent presenting a course on blood pressure, followed by a multiple-choice test. The results of this study suggest that the participants learned more from the male agent than from the female one. The researchers suggested that this might be because the female tutor did not conform to the stereotype of men as teachers. The first study showed that participants in this experiment perceived the female agent as very feminine, while the male agent was found to be very masculine. This may be due to a difference in the participants' interpretation of the female agent as being "too feminine" to be suitable for the role of tutor. The study did not address how participants perceived the agents' expertise or seek to determine any possible interactions between perceived expertise, the perceived agent's femininity, and performance on the test. Gender, while important, must be taken into account in combination with other features. For instance, Krämer et al. (2016) analyzed the impact of pedagogical agents' gender and their behavior on adults' motivation, effort, and performance in math. They found that the simple presence of a female virtual agent in a learning situation did not increase women's motivation and learning. However, when the agent displayed human-like non-verbal behavior by aligning with the participants' non-verbal behavior (Gratch et al., 2007), the participants' performance and effort improved. This kind of behavior, called rapport, is defined in social psychology as the establishment of a positive relationship between interactants by way of a positive attitude (e.g., acquiescence, smiles), mutual attention (e.g., mutual gaze), and coordination of behaviors (e.g., synchrony, mimicry) (Tickle-Degnen and Rosenthal, 1990). This research shows the importance of the pedagogical agents' behavior combined with their gender as providing a positive impact on academic outcomes. Agents' behavior is especially important as it could negatively impact learners' academic outcomes, as shown in an experiment by Chang et al. (2019) where a male "dominant" pedagogical agent impaired female participants' performance and recall in arithmetic problems, compared to a male "non-dominant" agent.

The research presented above highlights the importance of pedagogical agents' gender on learning. Different studies appear to yield contradictory results, on one hand, that learning improves when virtual agents' gender matches the learner's, but on the other hand that male virtual agents could be better suited to improving learning. Interestingly, Section 3.4 shows that male agents are perceived as more competent than female agents, and users follow more advice from a male agent than a female one on topics stereotypically perceived as masculine. However, the studies featuring a female agent in STEM fields (Science, Technology, Engineering, and Mathematics) presented in this section show that female agents have a positive influence on academic outcomes: they improve learning, self-efficacy, interest, and motivation, despite the fact that STEMs are perceived as masculine (Makarova et al., 2019).

4. DISCUSSION

4.1. The Question of Pedagogical Agents' Gender

Based on the research presented above, one could surmise that, in general, male pedagogical agents are better suited to improving academic outcomes than female agents. However, systematically relying on male pedagogical agents could have an adverse impact: for instance, designing only male agents for learning purposes in STEM fields could strengthen gender stereotypes. As highlighted by West et al. (2019), the gender bias of interactive systems not only perpetuates stereotypes but also reinforces and extends them. The stereotypes modeled through interactive systems generate behaviors that go beyond the sphere of the virtual environment by conveying a harmful image of women. For instance, in a study by Dill et al. (2008), still pictures of men and women in suits or male and female characters acting in highly stereotypical ways were shown to participants. Male participants exposed to negative female stereotypes were significantly more tolerant of a real-life instance of sexual harassment and exhibited greater rape myth acceptance. As for representation in STEM fields, as noted by Sansone (2019), the lack of female role models can lead female students to believe that men are better than women in STEM fields. The lack of *virtual* female role models in virtual learning environments may have the same impact. Accordingly, more STEM experts represented with virtual female characters could help decrease gender stereotypes in STEM fields.

Some research has explored the use of androgynous virtual agents to counter gender stereotypes. In earlier studies, participants tended to apply the labels of "man" or "woman" to androgynous agents. For instance, in Niculescu et al. (2009), participants classified androgynous virtual characters as male or female, depending on the participants' gender and other parameters such as which virtual characters they had seen before. Even for genderless agents such as a wooden mannequin, the participants perceived their gender depending on how they perceived their walking motions (McDonnell et al., 2009). Recent research has shown more promising results in terms of gender stereotypes. In Nag and Yalçın (2020), results for androgynous agents show a linear trend that positions their scores for the perceived agency, communion, and competence in between those for female and male agents. The authors, thus, believe that androgynous agents could help mitigate male and female stereotypes. Although participants in their first experiment tended to believe that the androgynous virtual agents were men, when the authors modified the agents in question for their main experiment, participants correctly perceived them as androgynous after reading a definition of an androgynous agent.

What about androgynous pedagogical agents in an educational context? Silvervarg et al. (2013) supposed, but with caution, that students could identify with an androgynous agent by ascribing their own gender to them, thus making them a suitable role model. Indeed, in their experiment with children aged 12-14, participants perceived an androgynous pedagogical agent as not clearly a boy nor clearly a girl, but they generally assigned themselves a gender to their androgynous virtual tutee, boy or girl. The authors supposed students could, therefore, have more freedom to construct and ascribe gender, as their pedagogical agent's gender choice is personal rather than imposed. They also supposed androgynous agents could diminish gender stereotypes, as their appearances are genderless. Applying our own gender to an androgynous agent to make them a suitable role model is an interesting hypothesis. However, we do not know what the gender participants applied to the androgynous agent or why. More research on androgynous agents has to be done in an educational context to help determine, e.g., whether androgynous agents are perceived as masculine, feminine, neutral, man, woman, or genderless depending on the context and the role of the agent. Since STEM fields are considered masculine fields (Makarova et al., 2019), participants could perceive an androgynous agent as a man, even though they could perceive them as not clearly a boy nor a girl in terms of appearance. This could reinforce the stereotype of STEM fields as more suitable for men than women. Agents' role is also particularly important, as Brahnam and Weaver (2015) stated there are more female assistant agents than male ones. They showed the example of a webpage that provides virtual agents, four of the five virtual agents are female and they assist people at airports or serve as talking mannequins for fashion and museum exhibits. The male agent was called a "virtual doctor" and provided health tips and hospital information. We can emit the hypothesis that one could perceive androgynous virtual assistants as women, hence reinforcing gender stereotypes. For this research on androgynous virtual agents, we recommend measuring how participants feel toward the androgynous agents, as not being able to perceive someone as a man or a woman may induce insecurity and unease in some people (Nass and Brave, 2005).

4.2. Virtual Agents' as Social Role Models in Learning Environments

Some research, though still very limited, has explored the use of virtual agents to increase learners' performance and interest in mathematics. For example, Rosenberg-Kima et al. (2008) showed the effectiveness of a female virtual agent engineer in interesting women in STEM fields. In a video, the agent, who was similar in gender to the participants (who were all women), presented a story about successful female role models in STEM fields. This led to a change in participants' attitudes toward science, as shown with a 7-point scale questionnaire. Women in the female virtual agent condition were less likely to endorse traditional STEM stereotypes than those in the male virtual agent condition and were more likely to believe that women could succeed in STEM fields. Gender stereotypes still persisted: the participants were slightly more likely to believe in STEM usefulness with a male virtual agent engineer. In a similar study by Plant et al. (2009), male and female participants performed better and were more interested in engineering after interacting with a female agent, as their self-efficacy and their ratings about STEM usefulness improved. Interestingly, male participants were less likely to endorse traditional STEM stereotypes in the presence of an agent, male or female; but female participants were less likely to endorse traditional STEM

stereotypes with a female agent or without any agent, than with a male one. Another similar study by Rosenberg-Kima et al. (2010) showed that Black virtual agents had a more positive impact on STEM interest and STEM gender stereotypes for Black women, whereas female virtual agents (Black or White) had a more positive impact on White women on the same criteria. This research shows the importance of other factors, such as virtual agents' ethnic background, performance, and interest in math.

Finally, several studies have shown that pedagogical agents used as learning companions can simulate social interactions (Kim and Baylor, 2006) and the potential impact of a virtual agent's gender on education. However, only few studies have explored the use of a virtual pedagogical companion to counteract the effects of Stereotype Threat (refer to Introduction). Research on Social Cognition has shown the positive impact of social role models to counteract ST effects (Bagès et al., 2016). Studies have shown that female participants do not immediately see female scientists as potential role models simply by interacting with them; they begin to perceive female scientists as role models when they establish personal connections with them (Buck et al., 2008). In the field of virtual agents, virtual rapport has been studied as a means to create this type of relationship between virtual agents and users (Gratch et al., 2007). As reported by Krämer et al. (2016), the mere presence of a female agent did not improve participants' performance and effort. However, when agents were able to create a virtual rapport, participants' performance and effort were shown to improve.

Based on the research presented above, not only is the gender of pedagogical agents important, but so is their behavior (Krämer et al., 2016; Chang et al., 2019), their role (Baylor and Kim, 2004; Kim, 2016), and their ethnicity (Rosenberg-Kima et al., 2010; Kim, 2016). Girls may see a female pedagogical agent as a role model who influences their motivation to exert effort to learn (Shiban et al., 2015). A study by Pfeifer and Lugrin (2018) shows that female social robots can be role models to female students: female students learned better with a female robot in a stereotypically masculine domain. Virtual characters can be used to embody social models and, thus, change the learner's attitudes and motivation; as described earlier, a female role model who succeeds in math can reduce Stereotype Threat effects. Combining research on social cognition and virtual agents, we recommend counteracting Stereotype Threat effects for girls and women in math by using a virtual agent representing a hardworking female social role model (Bagès et al., 2016) able to establish rapport with the learners (Gratch et al., 2007), of similar ethnicity to the learners (Kim, 2016) and slightly older than them (Bagès and Martinot, 2011). When the role model is younger or the same age as the learners, they can lose motivation by feeling unable to match their role model's achievements; if the model is too old, they will not identify with them. A pedagogical agent should, thus, embody the role of a knowledgeable and motivational person; this has been demonstrated by student preferences and by the proven positive impact these types of agents have on education (Kim and Baylor, 2016).

4.3. Improved Learning or Better Inclusion?

An ethical tension between two competing goals arises in all domains: skill learning (where the research presented above favors the use of a male virtual agent), vs. better inclusion of girls and women (*via* the use of a female virtual character embodying a successful role model in the domain). Prior research is not robust enough to prove the superiority of a male agent in all fields and for all audiences. Some questions remain unanswered in the literature, to our knowledge: Would using the same androgynous character but presented as male, female, or neutral by the experimenter have an impact on academic outcomes in scientific or other domains? What would be the impact of systematically using a virtual agent of the same gender as the learners?

Regarding the second question, using only successful male models in mathematics with boys could reinforce gender stereotypes. Women are aware of the negative stereotype about their mathematical skills that create a hostile environment for them. Research by Stokes et al. (1995) reported that when women find a friendly environment, they are more likely to stay employed. One solution to reconciling the two goals, at least in STEM fields, would be to use successful female role models to explain how they managed to perform well: in Bagès et al. (2016), students took a math test after reading a story about a social model, female or male depending on the condition. The stories differentiated between models: the hardworking model put in the effort and spent time learning his or her lessons to perform well, the gifted model was naturally good at math, and the neutral model gave no explanation for his or her success. Girls' performance increased with the hardworking model: they performed at the same level as boys, whether the model was a boy or a girl. There was no impact on boys' performance. In contrast, in a similar study, boys' performance also increased with a hardworking model, regardless of gender (Bagès and Martinot, 2011). Furthermore, when the role model did not explain his or her success in math, both girls' and boys' scores improved with a female role model. As the lack of female role models may lead female students to believe that men are better than women in STEM fields Sansone (2019), it could be interesting to combine the results of (Bagès et al., 2016) and our hypothesis that a successful *virtual* female role model in STEM fields could help mitigate gender stereotypes. Female virtual agents who act as successful social models in mathematics and explain how they succeeded through their effort and hard work may be a potential solution to counteracting Stereotype Threat effects. Another question then arises: in the long term, what would be the impact of presenting only these kinds of female virtual agents to boys?

5. CONCLUSION

In this article we have presented a systematic review of research on perceptive studies of virtual agents depending on their gender, regardless of the application domain; and on the impact of gendered virtual agents in the context of a learning task. Each study has been performed in a specific learning context

with specific pedagogy, design of the virtual environment, duration of the interaction, modality of interaction, physical environment, etc. These elements of context may have an impact on the users' perceptions and learning outcomes. The limitation of this article is that we have not considered all these contextual specificities. Further analysis could take into account these contextual elements to provide a more global view of the impact of virtual agents' gender in academic learning. Nevertheless, the present systematic review enables us to draw some conclusions by answering each question stated in Section 2.2.

Do Students Perceive Virtual Agents' Differently Depending on Their Own Gender and the Gender of the Agent?

As individuals communicate with virtual agents by applying social rules and expectations as social beings (Nass and Moon, 2000), it is not surprising that they also apply gender stereotypes to virtual agents and their interactions with them (refer to Section 3.4). Female virtual agents are usually seen as less expert, less knowledgeable, and less powerful than male virtual agents (Baylor and Kim, 2004; Nunamaker et al., 2011), and they are also usually perceived as more likable and attractive than male virtual agents (Nunamaker et al., 2011; Lunardo et al., 2016). Those perception differences can even affect people's decisions (Lee, 2003; Świdrak et al., 2021).

Given the empirical results, we propose to respond simultaneously to question 2 (*Does the gender of pedagogical agents influence students' academic performance and self-evaluations?*) and question 3 (*Are there tasks or academic situations to which a male virtual agent is better suited to than a female virtual agent, and vice versa, according to empirical evidence?*). The review conducted in this article could lead to the belief that male pedagogical agents are better suited than female agents to improve academic outcomes, especially in male-dominated scientific fields like STEM fields (Makarova et al., 2019). However, research also shows that female pedagogical agents can improve learners' performances in these fields (Rosenberg-Kima et al., 2008; Plant et al., 2009). Some studies have shown that using female virtual agents as social role models can increase female participants' self-efficacy (Rosenberg-Kima et al., 2008), and both male and female participants' interest (Plant et al., 2009; Rosenberg-Kima et al., 2010), and performances (Plant et al., 2009). These results are especially relevant to addressing Stereotype Threat effects, a phenomenon that illustrates how and why female students' performance in math can be impaired by gender stereotypes (Spencer et al., 1999). Previous research showed that stereotype threat effects can be counteracted by introducing a female positive role model who looks like a learner so that they can identify with her (Bagès et al., 2016). Such a positive role model could be embodied by a virtual pedagogical agent (Rosenberg-Kima et al., 2008), and used to reduce stereotype threats in STEM fields.

How Do a Virtual Agent's Pedagogical Roles Impact These Results?

Individuals tend to listen more to agents whose gender stereotypically matches the context or gender roles such as female virtual agents with cosmetics and male virtual agents with sports (Lee, 2003), female virtual agents in contexts involving social influence (Khashe et al., 2017), and female virtual agents in an assistant role (Esposito et al., 2021). The role of a virtual pedagogical agent has been studied in the academic context, using *expert*, *motivator*, and *mentor* agents (Baylor and Kim, 2004). Expert agents are older than students, authoritative, strictly informative, and knowledgeable. Motivator agents are enthusiastic and not seen as particularly knowledgeable, they are mostly used to elicit motivation. As for mentor agents, they are slightly older than students, are knowledgeable, and are also used to elicit motivation. They are a mix of expert agents and motivator agents (Baylor and Kim, 2005). Researchers should take agents' roles into account when designing a pedagogical agent, as a female pedagogical agent designing as a mentor agent can improve learners' performance (Plant et al., 2009), but a female pedagogical agent designing as a motivator agent may not be effective on learners' performance (Shiban et al., 2015). In conclusion, the impact of the virtual agent's gender depends on several aspects related to the role of the agent. This role is indeed reflected through for instance the appearance but also the discourse of the agent.

How Do a Virtual Agent'S Appearance and Interactive Capacities Impact These Results?

The topics of individuals' interactions with virtual agents differ depending on their gendered appearance: in studies by De Angeli and Brahnam (2006) and Brahnam and De Angeli (2012), female virtual agents received significantly more violent sexual propositions, more rape threats, more comments on their appearance, and more swear words compared to male virtual agents who received few sexual propositions, most of them targeting their girlfriends. Moreover, the degree of perceived masculinity and femininity can influence men's decisions, as shown in Świdrak et al. (2021) where male participants were persuaded more by masculine agents than feminine agents, regardless of the agents' gender. Agents' appearance can influence their perceived role, as an old agent wearing a tie could be perceived more as an expert than a young agent (Shiban et al., 2015). As seen in the question above, the agents' role is important in academic situations. In addition to agents' roles and appearance, research has shown the importance of a positive relationship between learners and pedagogical agents (Krämer et al., 2016). This research tends to show that a female social model embodied by a pedagogical agent able to establish a positive relationship with learners may counteract Stereotype Threat effects and, thus, improve women's performance, interest, and self-efficacy in mathematics.

Are Androgynous Virtual Agents' a Potential Solution to Combatting Gender Stereotypes?

This question is quite difficult to answer as to our knowledge, few studies have explored the use of androgynous virtual agents to counter gender stereotypes in the academic context. Silvervarg et al. (2013) used an androgynous pedagogical agent in an educational context and showed participants, although evaluating the agent as "not clearly a boy nor a girl," tend to ascribe a binary gender (boy or girl) to the agent. The authors, thus, cautiously supposed that students could ascribe their own gender to an androgynous agent, thus giving them more freedom and making the agent a suitable role model, known to be beneficial for academic outcomes. However, the results did not show what gender participants ascribe to the agent, nor why. The way individuals ascribe gender to an androgynous or a genderless agent should be studied more. In particular, does this gender attribution depend on the context or agents' role? Since STEM fields are considered masculine fields (Makarova et al., 2019), participants could think androgynous agents are men. This could reinforce the stereotype of STEM fields as more suitable for men than women. As for agents' roles, there are more female virtual assistants than male ones (Brahnam and Weaver, 2015). Some developers admitted female virtual assistants are usually used because they evoke gender stereotypes: women are expected to serve, help, and nurture others. Androgynous virtual assistants could then be considered women, and reinforce harmful stereotypes about women. Researchers and developers who want to use androgynous agents to combat gender stereotypes should be very careful, as the opposite effect can occur.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

MA wrote the article. MO and IR supervised the article and corrected it. All authors contributed to the article and approved the submitted version.

FUNDING

This work was carried out within the pilot center Ampiric, funded by the French State's Future Investment Program (PIA3/France 2030) as part of the "Territories of Educational Innovation" action.

REFERENCES

Ait Challal, T., and Grynszpan, O. (2018). "What gaze tells us about personality," in *Proceedings of the 6th International Conference on Human-Agent Interaction* (Southampton), 129–137.

Akbar, F., Grover, T., Mark, G., and Zhou, M. X. (2018). "The effects of virtual agents' characteristics on user impressions and language use, in *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (Tokyo), 1–2.

Antonio Gómez-Jáuregui, D., Giraud, T., Isableu, B., and Martin, J.-C. (2021). Design and evaluation of postural interactions between users and a listening virtual agent during a simulated job interview. *Comput. Animat. Virtual Worlds* 22, e2029. doi: 10.1002/cav.2029

Astrid, M., Krämer, N. C., and Gratch, J. (2010). "How our personality shapes our interactions with virtual characters-implications for research and development," in *International Conference on Intelligent Virtual Agents* (Philadelphia, PA: Springer), 208–221.

Bagès, C., and Martinot, D. (2011). What is the best model for girls and boys faced with a standardized mathematics evaluation situation: a hardworking role model or a gifted role model? *Br. J. Soc. Psychol.* 50, 536–543. doi: 10.1111/j.2044-8309.2010.02017.x

Bagès, C., Verniers, C., and Martinot, D. (2016). Virtues of a hardworking role model to improve girls" mathematics performance. *Psychol. Women Q.* 40, 55–64. doi: 10.1177/0361684315608842

Bailenson, J. N., Blascovich, J., Beall, A. C., and Loomis, J. M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence* 10, 583–598. doi: 10.1162/105474601753272844

Bandura, A., and National Inst of Mental Health. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Princeton, NJ: Prentice-Hall, Inc.

Baylor, A. L., and Kim, Y. (2004). "Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role," in *7th International Conference in Intelligent Tutoring Systems* (Maceió-Alagoas).

Baylor, A. L., and Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *Int. J. Artif. Intell. Educ.* 15, 95.

Bem, S. L. (1974). The measurement of psychological androgyny. *J. Consult. Clin. Psychol.* 42, 155. doi: 10.1037/h0036215

Biocca, F. (1997). The cyborg's dilemma: progressive embodiment in virtual environments. *J. Comput. Mediat. Commun.* 3, JCMC324. doi: 10.1111/j.1083-6101.1997.tb00070.x

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: review and suggested criteria. *Presence* 12, 456–480. doi: 10.1162/105474603322761270

Biocca, F., Harms, C., and Gregg, J. (2001). "The networked minds measure of social presence: pilot test of the factor structure and concurrent validity," in *4th Annual International Workshop on Presence* (Philadelphia, PA), 1–9.

Brahnam, S., and De Angeli, A. (2012). Gender affordances of conversational agents. *Interact. Comput.* 24, 139–153. doi: 10.1016/j.intcom.2012.05.001

Brahnam, S., and Weaver, M. (2015). "Re/framing virtual conversational partners: a feminist critique and tentative move towards a new design paradigm," in *International Conference of Design, User Experience, and Usability* (Los Angeles, CA: Springer), 172–183.

Buck, G. A., Clark, V. L. P., Leslie-Pelecky, D., Lu, Y., and Cerda-Lizarraga, P. (2008). Examining the cognitive processes used by adolescent girls and women scientists in identifying science role models: a feminist approach. *Sci. Educ.* 92, 688–707. doi: 10.1002/sce.20257

Chang, F., Luo, M., Walton, G., Aguilar, L., and Bailenson, J. (2019). Stereotype threat in virtual learning environments: effects of avatar gender and sexist behavior on women's math learning outcomes. *Cyberpsychol. Behav. Soc. Network.* 22, 634–640. doi: 10.1089/cyber.2019.0106

Cloud-Buckner, J., Sellick, M., Sainathuni, B., Yang, B., and Gallimore, J. (2009). "Expression of personality through avatars: analysis of effects of gender and race on perceptions of personality," in *International Conference on Human-Computer Interaction* (Uppsala: Springer), 248–256.

De Angeli, A., and Brahnam, S. (2006). "Sex stereotypes and conversational agents," in *Proceedings of Gender and Interaction: Real and Virtual Women in a Male World* (Venice).

Dill, K. E., Brown, B. P., and Collins, M. A. (2008). Effects of exposure to sex-stereotyped video game characters on tolerance of sexual harassment. *J. Exp. Soc. Psychol.* 44, 1402–1408. doi: 10.1016/j.jesp.2008.06.002

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., and Sczesny, S. (2020). Gender stereotypes have changed: a cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *Am. Psychol.* 75, 301. doi: 10.1037/amp0000494

Ellis, A. (1962). *Reason and Emotion in Psychotherapy*. Secaucus, NJ: Citadel Press.

Esposito, A., Amorese, T., Cuciniello, M., Riviello, M. T., Esposito, A. M., Troncone, A., et al. (2021). Elder user's attitude toward assistive virtual agents: the role of voice and gender. *J. Ambient. Intell. Humaniz Comput.* 12, 4429–4436. doi: 10.1007/s12652-019-01423-x

Feng, D., Jeong, D. C., Krämer, N. C., Miller, L. C., and Marsella, S. (2017). "is it just me?": evaluating attribution of negative feedback as a function of virtual instructor's gender and proxemics," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent System* (São Paulo), 810–818.

Forgasz, H. J., and Leder, G. C. (1996). Mathematics classrooms, gender and affect. *Math. Educ. Res. J.* 8, 153–173. doi: 10.1007/BF03217295

Fox, J., and Bailenson, J. N. (2009). Virtual virgins and vamps: the effects of exposure to female characters" sexualized appearance and gaze in an immersive virtual environment. *Sex Roles* 61, 147–157. doi: 10.1007/s11199-009-9599-3

Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). "Creating rapport with virtual agents," in *International Workshop on Intelligent Virtual Agents* (Stockholm: Springer), 125–138.

Guadagno, R. E., Blascovich, J., Bailenson, J. N., and McCall, C. (2007). Virtual humans and persuasion: the effects of agency and behavioral realism. *Media Psychol.* 10, 1–22. doi: 10.108/15213260701300865

Guay, F., Vallerand, R. J., and Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: the Situational Motivation Scale (SIMS). *Motiv. Emot.* 24, 175–213. doi: 10.1023/A:1005614228250

Gulz, A., Haake, M., and Tärning, B. (2007). Visual gender and its motivational and cognitive effects-a user study. *Lund Univers. Cogn. Stud.* 137, 928–935.

Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Adv. Psychol.* 52, 139–183. doi: 10.1016/S0166-4115(08)62386-9

Hayes, A. L., Ulinski, A. C., and Hodges, L. F. (2010). "That avatar is looking at me! Social inhibition in virtual worlds," in *International Conference on Intelligent Virtual Agents* (Philadelphia, PA: Springer), 454–467.

Jeong, D. C., Feng, D., Krämer, N. C., Miller, L. C., and Marsella, S. (2017). "Negative feedback in your face: examining the effects of proxemics and gender on learning," in *International Conference on Intelligent Virtual Agents* (Stockholm: Springer), 170–183.

Johnson, A. M., Ozogul, G., Moreno, R., and Reisslein, M. (2013). Pedagogical agent signaling of multiple visual engineering representations: the case of the young female agent. *J. Eng. Educ.* 102, 319–337. doi: 10.1002/jee.20009

Kantharaju, R. B., De Franco, D., Pease, A., and Pelachaud, C. (2018). "Is two better than one? Effects of multiple agents on user persuasion," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW), 255–262.

Khashe, S., Lucas, G., Becerik-Gerber, B., and Gratch, J. (2017). Buildings with persona: towards effective building-occupant communication. *Comput. Hum. Behav.* 75, 607–618. doi: 10.1016/j.chb.2017.05.040

Kim, Y. (2013). Digital peers to help children's text comprehension and perceptions. *J. Educ. Technol. Soc.* 16, 59–70.

Kim, Y. (2016). The role of agent age and gender for middle-grade girls. *Comput. Sch.* 33, 59–70. doi: 10.1080/07380569.2016.1143753

Kim, Y., and Baylor, A. L. (2006). A social-cognitive framework for pedagogical agents as learning companions. *Educ. Technol. Res. Dev.* 54, 569–596. doi: 10.1007/s11423-006-0637-3

Kim, Y., and Baylor, A. L. (2016). Research-based design of pedagogical agent roles: A review, progress, and recommendations. *Int. J. Artif. Intell. Educ.* 26, 160–169. doi: 10.1007/s40593-015-0055-y

Kim, Y., Baylor, A. L., and Shen, E. (2007). Pedagogical agents as learning companions: the impact of agent emotion and gender. *J. Comput. Assist. Learn.* 23, 220–234. doi: 10.1111/j.1365-2729.2006.00210.x

Kim, Y., and Lim, J. H. (2013). Gendered socialization with an embodied agent: creating a social and affable mathematics learning environment for middle-grade females. *J. Educ. Psychol.* 105, 1164. doi: 10.1037/a0031027

Kim, Y., and Wei, Q. (2011). The impact of learner attributes and learner choice in an agent-based environment. *Comput. Educ.* 56, 505–514. doi: 10.1016/j.compedu.2010.09.016

Krämer, N. C., Karacora, B., Lucas, G., Dehghani, M., Rüther, G., and Gratch, J. (2016). Closing the gender gap in with friendly male instructors? On the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Comput. Educ.* 99, 1–13. doi: 10.1016/j.compedu.2016.04.002

Kulms, P., Krämer, N. C., Gratch, J., and Kang, S.-H. (2011). "It's in their eyes: a study on female and male virtual humans' gaze," in *International Workshop on Intelligent Virtual Agents* (Reykjavik: Springer), 80–92.

Lee, E.-J. (2003). Effects of "gender" of the computer on informational social influence: the moderating role of task type. *Int. J. Hum. Comput. Stud.* 58, 347–362. doi: 10.1016/S1071-5819(03)00009-0

Li, J., Kizilcec, R., Bailenson, J., and Ju, W. (2016). Social robots and virtual agents as lecturers for video instruction. *Comput. Hum. Behav.* 55, 1222–1230. doi: 10.1016/j.chb.2015.04.005

Lindberg, S. M., Hyde, J., Petersen, J., and Linn, M. (2010). Gender similarities characterize math performance. *Psychol. Bull.* 136:1123–1135. doi: 10.1037/a0021276

Lunardo, R., Bressolles, G., et al. (2016). The interacting effect of virtual agents' gender and dressing style on attractiveness and subsequent consumer online behavior. *J. Retail. Consum. Serv.* 30, 59–66. doi: 10.1016/j.jretconser.2016.01.006

Makarova, E., Aeschlimann, B., and Herzog, W. (2019). The gender gap in fields: the impact of the gender stereotype of math and science on secondary students' career aspirations. *Front. Educ.* 4, 60. doi: 10.3389/feduc.2019.00060

Makransky, G., Wismer, P., and Mayer, R. E. (2019). A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. *J. Comput. Assist. Learn.* 35, 349–358. doi: 10.1111/jcal.12335

McDonnell, R., Jörg, S., Hodgins, J. K., Newell, F., and O'Sullivan, C. (2009). Evaluating the effect of motion and body shape on the perceived sex of virtual characters. *ACM Trans. Appl. Percept.* 5, 1–14. doi: 10.1145/1462048.1462051

Mell, J., Lucas, G., and Gratch, J. (2017). "Prestige questions, online agents, and gender-driven differences in disclosure," in *International Conference on Intelligent Virtual Agents* (Stockholm: Springer), 273–282.

Mohtadi, M. T., Hajami, A., and Allali, H. (2014). "Pedagogical agent for metacognitive scaffolding in interactive learning environments," in *2014 International Conference on Multimedia Computing and Systems (ICMCS)* (Marrakech: IEEE), 652–656.

Moreno, K. N., Person, N. K., Adcock, A. B., Eck, R., Jackson, G. T., and Marineau, J. C. (2002). "Etiquette and efficacy in animated pedagogical agents: the role of stereotypes," in *AAAI Symposium on Personalized Agents* (Cape Cod, MA).

Moreno, R., and Flowerday, T. (2006). Students' choice of animated pedagogical agents in science learning: a test of the similarity-attraction hypothesis on gender and ethnicity. *Contemp Educ. Psychol.* 31, 186–207. doi: 10.1016/j.cedpsych.2005.05.002

Mousas, C., Anastasiou, D., and Spantidi, O. (2018). The effects of appearance and motion of virtual characters on emotional reactivity. *Comput. Hum. Behav.* 86, 99–108. doi: 10.1016/j.chb.2018.04.036

Nag, P., and Yalçın, Ö. N. (2020). "Gender stereotypes in virtual agents," in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Glasgow), 1–8.

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Nass, C. I., and Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA: MIT Press.

Niculescu, A., Hofs, D., Van Dijk, B., and Nijholt, A. (2010). "How the agent's gender influence users' evaluation of a qa sy," in *2010 International Conference on User Science and Engineering (i-USEr)* (Shah Alam, Selangor: IEEE), 16–20.

Niculescu, A., Van Der Sluis, F., and Nijholt, A. (2009). "Feminity, masculinity and androgyny: how humans perceive the gender of anthropomorphic agents," in *Proceedings of 13th International Conference on Human-Computer Interaction* (Heidelberg: Springer Verlag), 628–632.

Nunamaker, J. F., Derrick, D. C., Elkins, A. C., Burgoon, J. K., and Patton, M. W. (2011). Embodied conversational agent-based kiosk for automated interviewing. *J. Manag. Inf. Syst.* 28, 17–48. doi: 10.2753/MIS0742-1222280102

Ozogul, G., Johnson, A. M., Atkinson, R. K., and Reisslein, M. (2013). Investigating the impact of pedagogical agent gender matching and learner choice on learning outcomes and perceptions. *Comput. Educ.* 67, 36–50. doi: 10.1016/j.compedu.2013.02.006

Pansu, P., Régner, I., Max, S., Colé, P., Nezlek, J. B., and Huguet, P. (2016). A burden for the boys: evidence of stereotype threat in boys' reading performance. *J. Exp. Soc. Psychol.* 65:26–30. doi: 10.1016/j.jesp.2016.02.008

Payne, J., Szymkowiak, A., Robertson, P., and Johnson, G. (2013). "Gendering the machine: Preferred virtual assistant gender and realism in self-service," in *International Workshop on Intelligent Virtual Agents* (Edinburgh: Springer), 106–115.

Pecune, F., Cafaro, A., Ochs, M., and Pelachaud, C. (2016). "Evaluating social attitudes of a virtual tutor," in *International Conference on Intelligent Virtual Agents* (Los Angeles, CA: Springer), 245–255.

Pezzullo, L. G., Wiggins, J. B., Frankosky, M. H., Min, W., Boyer, K. E., Mott, B. W., et al. (2017). "Thanks alisha, keep in touch": gender effects and engagement with virtual learning companions," in *International Conference on Artificial Intelligence in Education* (Wuhan: Springer), 299–310.

Pfeifer, A., and Lugrin, B. (2018). "Female robots as role-models? The influence of robot gender and learning materials on learning success," in *International Conference on Artificial Intelligence in Education* (London: Springer), 276–280.

Plant, E. A., Baylor, A. L., Doerr, C. E., and Rosenberg-Kima, R. B. (2009). Changing middle-school students' attitudes and performance regarding engineering with computer-based social models. *Comput. Educ.* 53, 209–215. doi: 10.1016/j.compedu.2009.01.013

Räty, H., Vänskä, J., Kasanen, K., and Kärkkäinen, R. (2002). Parents' explanations of their child's performance in mathematics and reading: a replication and extension of yee and eccles. *Sex Roles* 46, 121–128. doi: 10.1023/A:1016573627828

Régner, I., Steele, J. R., Ambady, N., Thinus-Blanc, C., and Huguet, P. (2014). Our future scientists: a review of stereotype threat in girls from early elementary school to middle school. *Revue Int. Psychol. Soc.* 27, 13–51.

Richards, D., Alsharbi, B., and Abdulrahman, A. (2020). "Can I help you? Preferences of young adults for the age, gender and ethnicity of a virtual support person based on individual differences including personality and psychological state," in *Proceedings of the Australasian Computer Science Week Multiconference* (Melbourne, VIC), 1–10.

Rosenberg-Kima, R., Baylor, A., Plant, E., and Doerr, C. (2008). Interface agents as social models for female students: the effects of agent visual presence and appearance on female students' attitudes and beliefs. *Comput. Hum. Behav.* 24, 2741–2756. doi: 10.1016/j.chb.2008.03.017

Rosenberg-Kima, R. B., Plant, E. A., Doerr, C. E., and Baylor, A. L. (2010). The influence of computer-based model's race and gender on female students' attitudes and beliefs towards engineering. *J. Eng. Educ.* 99, 35–44. doi: 10.1002/j.2168-9830.2010.tb01040.x

Sajjadi, P., Zhao, J., Wallgrün, J. O., Furman, T., La Femina, P. C., Fatemi, A., et al. (2020). "The effect of virtual agent gender and embodiment on the experiences and performance of students in virtual field trips," in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (Takamatsu: IEEE), 221–228.

Sansone, D. (2019). Teacher characteristics, student beliefs, and the gender gap in STEM fields. *Educ. Eval. Policy Anal.* 41, 127–144. doi: 10.3102/0162373718819830

Schmader, T., and Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *J. Pers. Soc. Psychol.* 85, 440. doi: 10.1037/0022-3514.85.3.440

Schroeder, N. L., Adesope, O. O., and Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *J. Educ. Comput. Res.* 49, 1–39. doi: 10.2190/EC.49.1.a

Schroeder, N. L., Romine, W. L., and Craig, S. D. (2017). Measuring pedagogical agent persona and the influence of agent persona on learning. *Comput. Educ.* 109, 176–186. doi: 10.1016/j.compedu.2017.02.015

Sczesny, S., Nater, C., and Eagly, A. H. (2018). "Agency and communion: Their implications for gender stereotypes and gender identities," in *Agency and Communion in Social Psychology* (New York, NY: Routledge), 103–116.

Shiban, Y., Schelhorn, I., Jobst, V., Hörnlein, A., Puppe, F., Pauli, P., et al. (2015). The appearance effect: influences of virtual agent features on performance and motivation. *Comput. Hum. Behav.* 49, 5–11. doi: 10.1016/j.chb.2015.01.077

Silvervarg, A., Haake, M., and Gulz, A. (2013). "Educational potentials in visually androgynous pedagogical agents," in *International Conference on Artificial Intelligence in Education* (Memphis, TN: Springer), 599–602.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype threat and women's math performance. *J. Exp. Soc. Psychol.* 35, 4–28. doi: 10.1006/jesp.1998.1373

Spilioto-poulos, D., Makri, E., Vassilakis, C., and Margaris, D. (2020). Multimodal interaction: correlates of learners' metacognitive skill training negotiation experience. *Information* 11, 381. doi: 10.3390/info11080381

Steele, C. M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *J. Pers. Soc. Psychol.* 69, 797. doi: 10.1037/0022-3514.69.5.797

Stokes, J., Riger, S., and Sullivan, M. (1995). Measuring perceptions of the working environment for women in corporate settings. *Psychol. Women Q.* 19, 533–549. doi: 10.1111/j.1471-6402.1995.tb00091.x

Świdrak, J., Pochwatko, G., and Insabato, A. (2021). Does an agent's touch always matter? Study on virtual midas touch, masculinity, social status, and compliance in Polish men. *J. Multimodal User Interf.* 15, 163–174. doi: 10.1007/s12193-020-00351-x

ter Stal, S., Tabak, M., op den Akker, H., Beinema, T., and Hermens, H. (2020). Who do you prefer? The effect of age, gender and role on users' first impressions of embodied conversational agents in ehealth. *Int. J. Hum. Comput. Interact.* 36, 881–892. doi: 10.1080/10447318.2019.1699744

Tickle-Degnen, L., and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychol. Inq.* 1, 285–293. doi: 10.1207/s15327965pli0104_1

van der Lubbe, L. M., and Bosse, T. (2017). "Studying gender bias and social backlash via simulated negotiations with virtual agents," in *International Conference on Intelligent Virtual Agents* (Stockholm: Springer), 455–458.

Vilaro, M. J., Wilson-Howard, D. S., Neil, J. M., Tavassoli, F., Zalake, M. S., Lok, B. C., et al. (2021). A subjective culture approach to cancer prevention: rural black and white adults' perceptions of using virtual health assistants to promote colorectal cancer screening. *Health Commun.* doi: 10.1080/10410236.2021.1910166

Webster, J., and Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Q.* 26, xiii-xxiii. doi: 10.2307/4132319

West, M., Kraut, R., and Ei Chew, H. (2019). *I'd blush if I could: closing gender divides in digital skills through education*. Available online at: https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

Wirzberger, M., Schmidt, R., Georgi, M., Hardt, W., Brunnett, G., and Rey, G. D. (2019). Effects of system response delays on elderly humans cognitive performance in a virtual training scenario. *Scientific Rep.* 9, 1–12. doi: 10.1038/s41598-019-44718-x

Yee, D. K., and Eccles, J. S. (1988). Parent perceptions and attributions for children's math achievement. *Sex Roles* 19, 317–333. doi: 10.1007/BF00289840

Zanbaka, C., Goolkasian, P., and Hodges, L. (2006). "Can a virtual cat persuade you? The role of gender and realism in speaker persuasiveness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, QC), 1153–1162.

Zibrek, K., Niay, B., Olivier, A.-H., Hoyet, L., Pettré, J., and Mcdonnell, R. (2020). The effect of gender and attractiveness of motion on proximity in virtual reality. *ACM Trans. Appl. Percept.* 17, 1–15. doi: 10.1145/3419985

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Artificial Intelligence

TYPE Correction
PUBLISHED 11 October 2023
DOI 10.3389/frai.2023.1302277

OPEN ACCESS

APPROVED BY
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

*CORRESPONDENCE
Magalie Ochs
✉ magalie.ochs@lis-lab.fr
Isabelle Régner
✉ isabelle.regner@univ-amu.fr

RECEIVED 26 September 2023
ACCEPTED 27 September 2023
PUBLISHED 11 October 2023

CITATION
Armando M, Ochs M and Régner I (2023) Corrigendum: The impact of pedagogical agents' gender on academic learning: a systematic review. *Front. Artif. Intell.* 6:1302277. doi: 10.3389/frai.2023.1302277

Corrigendum: The impact of pedagogical agents' gender on academic learning: a systematic review

Marjorie Armando[1,2,3], Magalie Ochs[1]* and Isabelle Régner[2]*

[1]Aix Marseille Univ, CNRS, LIS UMR 7020, Marseille, France, [2]Aix Marseille Univ, CNRS, LPC, Marseille, France, [3]Pôle pilote Ampiric, Institut National Supérieur du Professorat et de l'Éducation, Aix-Marseille Université, Marseille, France

KEYWORDS

virtual agent, gender, pedagogical agent, learning environment, gender stereotypes, systematic review

A corrigendum on
The impact of pedagogical agents' gender on academic learning: a systematic review

by Armando, M., Ochs, M., and Régner, I. (2022). *Front. Artif. Intell.* 5:862997. doi: 10.3389/frai.2022.862997

In the published article, there was an error in affiliations 1 and 3. Instead of "Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France," it should be "Aix Marseille Univ, CNRS, LIS UMR 7020, Marseille, France." Instead of "Pôle pilôte Ampiric, Institut National Supérieur du Proffessorat et de l'Éducation, Aix-Marseille Université, Marseille, France," it should be "Pôle pilote Ampiric, Institut National Supérieur du Professorat et de l'Éducation, Aix-Marseille Université, Marseille, France."

In the published article, there was also an error in the Funding section, the Funding statement was not included. The correct Funding statement appears below.

Funding

This work was carried out within the pilot center Ampiric, funded by the French State's Future Investment Program (PIA3/France 2030) as part of the "Territories of Educational Innovation" action.

The authors apologize for these errors and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

Publisher's note

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 06 October 2022
DOI 10.3389/frai.2022.1006173

OPEN ACCESS

EDITED BY
Rashid Mehmood,
King Abdulaziz University, Saudi Arabia

REVIEWED BY
Ben Chester Cheong,
Singapore University of Social Sciences, Singapore
Chuan Chen,
University of Miami, United States

*CORRESPONDENCE
Lingnan He
heln3@mail.sysu.edu.cn

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 04 August 2022
ACCEPTED 20 September 2022
PUBLISHED 06 October 2022

CITATION
Wang H, Sun Q, Gu L, Lai K and He L (2022) Diversity in people's reluctance to use medical artificial intelligence: Identifying subgroups through latent profile analysis.
Front. Artif. Intell. 5:1006173.
doi: 10.3389/frai.2022.1006173

Diversity in people's reluctance to use medical artificial intelligence: Identifying subgroups through latent profile analysis

Haixia Wang[1], Qiaoqiao Sun[2], Li Gu[3], Kaisheng Lai[1] and Lingnan He[4]*

[1]School of Journalism and Communication, Jinan University, Guangzhou, China, [2]Guangdong Medical Doctor Association, Guangzhou, China, [3]School of Innovation Design, Guangzhou Academy of Fine Arts, Guangzhou, China, [4]School of Communication and Design, Sun Yat-sen University, Guangzhou, China

Medical artificial intelligence (AI) is important for future health care systems. Research on medical AI has examined people's reluctance to use medical AI from the knowledge, attitude, and behavioral levels in isolation using a variable-centered approach while overlooking the possibility that there are subpopulations of people who may differ in their combined level of knowledge, attitude and behavior. To address this gap in the literature, we adopt a person-centered approach employing latent profile analysis to consider people's medical AI objective knowledge, subjective knowledge, negative attitudes and behavioral intentions. Across two studies, we identified three distinct medical AI profiles that systemically varied according to people's trust in and perceived risk imposed by medical AI. Our results revealed new insights into the nature of people's reluctance to use medical AI and how individuals with different profiles may characteristically have distinct knowledge, attitudes and behaviors regarding medical AI.

KEYWORDS

medical AI, objective knowledge, subjective knowledge, negative attitude, behavioral intention

Introduction

Medical artificial intelligence (AI) is critical to the future of medical diagnosis and can provide expert-level medical decisions. For example, in telemedicine, it is crucial to apply medical AI for diagnoses such as COVID-19 and skin cancer (Esteva et al., 2017; Hao, 2020; Hollander and Carr, 2020; Wosik et al., 2020). This advantage is particularly critical for improving the level of medical care in poor areas of developing countries (Topol, 2019). Despite this importance, there are many barriers to applying medical AI in health-care systems (Dwivedi et al., 2021). A multitude of studies have documented these barriers, including the public not having enough AI knowledge and people expressing negative attitudes toward medical AI in social media (Promberger and Baron, 2006; Eastwood et al., 2012; Price, 2018; Cadario et al., 2021).

At the behavioral level, health-care system providers are reluctant to use medical AI, and patients hold doubts about using medical AI (Longoni et al., 2019). In light of previous studies on medical AI, it is critical for scholars to develop a better holistic understanding of how knowledge, negative attitudes and behavior factors are combined to influence the acceptance of medical AI by the population.

Thus far, the most common method to explore the obstacles in the application of medical AI is to ask people to self-report variables regarding their knowledge, attitude, and behavior toward AI and then to explore the relationships among these variables by using regression-based statistical analyses (Xu and Yu, 2019; Abdullah and Fakieh, 2020; Cadario et al., 2021). This approach represents a variable-centered method in which the unique relationships of each factor with other variables are explored (Marsh et al., 2009). However, such an approach does not reveal the ways in which individuals may have knowledge, negative attitudes and behavior factors that combine to shape their profile (Ekehammar and Akrami, 2003). For example, some individuals may have high knowledge while still having high negative attitudes toward medical AI. These ideas suggest that distinct profiles of medical AI likely exist. To investigate such a possibility, a person-centered approach is needed to explore the presence of distinct subpopulations of medical AI that differentially combine knowledge, negative attitudes and behavior (Zyphur, 2009; Wang and Hanges, 2011). Unfortunately, this approach to medical AI has mostly been overlooked. A person-centered approach allows researchers to understand how knowledge of and negative attitudes and behaviors toward medical AI conjointly shape profiles by capturing unobserved heterogeneity in the way people report their knowledge, negative attitudes and behaviors toward medical AI. These profiles can be leveraged to understand the barriers and further aid the application of medical AI. For example, the profile of low knowledge of but high negative attitude toward medical AI might be used to identify public policy to reduce the negative attitude toward medical AI by increasing the science knowledge of medical AI. Overall, there is value in examining whether there exist different profiles of barriers to medical AI.

To address these questions, we adopt the knowledge, attitudes and behavior (KAB) model (Kemm and Close, 1995; Yi and Hohashi, 2018) to understand the barriers to medical AI. The KAB model is particularly helpful and relevant for understanding and explaining the barriers to adopting medical AI. The core tenet of this model is that knowledge, attitudes, and behaviors are three related factors that are used to promote technology diffusion (Hohashi and Honda, 2015). Importantly, this model recognizes that these three factors are useful at identifying barriers to technology. Moreover, scholars have identified that the distinction between subjective knowledge and objective knowledge is important to understanding the barriers to medical AI. For instance, one recent study found that subjective knowledge of medical AI drives healthcare provider utilization (Cadario et al., 2021). Moreover, they found that greater subjective knowledge of medical decisions made by humans than medical AI providers contributes to medical AI aversion. Their findings imply how reluctance to utilize medical AI is driven both by the difficulty of subjectively understanding how medical AI makes decisions and by their objective understanding of human decision making. Drawing upon the KAB model, we investigate the profiles of heterogeneity in medical AI's objective knowledge, subjective knowledge, negative attitudes, and behavior.

Therefore, the objective of this research was to identify and describe the diversity in people's reluctance to use medical AI and its associated antecedents by employing latent profile analysis (LPA) (Woo et al., 2018). Specifically, we first tried to establish KAB profiles of medical AI in Study 1. Then, we sought to replicate and theoretically develop KAB profiles of medical AI in Study 2. Moreover, we tried to theoretically develop the KAB profiles by addressing the antecedents.

Study 1: Establishing KAB profiles of medical AI

In Study 1, we use an inductive approach to establish profiles of medical AI (Woo and Allen, 2014). A person-centered approach can establish quantitatively distinct profiles that differ in the levels of objectivity and knowledge of and negative attitudes and behaviors toward medical AI; it can also create qualitatively distinct profiles varying in the relative degree of objective knowledge and subjective knowledge of negative attitudes and behaviors toward medical AI. For instance, one profile may include people with high objective and subjective knowledge of as well as negative attitudes and behavior toward medical AI, while another includes low levels of objective and subjective knowledge of as well as negative attitudes and behavior toward medical AI. Given the various combinations that may occur, we pose the following question:

> *Research question: Are there distinct profiles of objective and subjective knowledge of and negative attitudes toward and behavior toward medical AI?*

Study 1: Methods

Participants and procedure

We recruited 328 participants online using convenience sampling. No participants were excluded. The participants provided informed consent and completed the survey. Table 1 provides demographic information on our sample.

TABLE 1 Demographic characteristics ($N = 328$).

Variables	Frequencies (percentages)
Age	
Mean (SD)	29.77 (8.18)
Sex	
Male	125 (38.1%)
Female	203 (61.9%)
Education year	
Mean (SD)	15.61 (1.75)
Occupation	
Full-time student	45 (13.7%)
Production	18 (5.5%)
Sales	23 (7.0%)
Public relations	19 (5.8%)
Customer service	11 (3.4%)
Administration	28 (8.5%)
Human resources	12 (3.7%)
Finance	21 (6.4%)
Clerical work	13 (4.0%)
Research	40 (12.2%)
Management	30 (9.1%)
Teaching	21 (6.4%)
Consulting	6 (1.8%)
Professional services (e.g., journalism and law)	22 (6.7%)
Other	19 (5.8%)

Measures

Objective knowledge of medical AI

We used the Cadario et al. (2021) three-item multiple choice test to measure the participants' objective understanding of medical AI. Each item had one correct answer for medical AI. We scored objective knowledge of medical AI by summing the correct answers. Thus, the objective knowledge of medical AI ranged from 0 to 3 ($m = 1.12$, $SD = 0.83$). Before the formal measurement, we interviewed doctors to ensure the accuracy of objective knowledge and expert validity.

Subjective knowledge of medical AI

We used the Cadario et al. (2021) three-item scale to measure the participants' subjective knowledge of medical AI. The participants were asked to indicate the extent to which they agreed with the included statements (1 = "don't quite understand," 5 = "quite understand"). One sample item is "To what extent do you feel that you understand what a medical AI algorithm considers when making the medical decision" ($\alpha = 0.74$).

Negative attitudes of medical AI

We measured negative attitudes toward medical AI using an 8-item scale (Schepman and Rodway, 2020). The participants were asked to indicate their level of agreement with a list of statements (1 = "Strongly disagree," 5 = "Strongly agree"). A sample item is as follows: "I find medical Artificial Intelligence sinister" ($\alpha = 0.84$).

Behavioral intention of medical AI use

We measured the behavioral intention of medical AI use using a 5-item scale (Esmaeilzadeh, 2020). The participants were asked to indicate their level of agreement with a list of statements (1 = "Strongly disagree," 5 = "Strongly agree"). A sample item is as follows: "I would like to use medical AI-based devices to manage my healthcare" ($\alpha = 0.84$).

Analytical approach

We first transformed raw measures of objective knowledge of medical AI, subjective knowledge of medical AI, negative attitudes toward medical AI, and behavioral intention toward medical AI use into z scores. Then, LPA was used to establish profiles of medical AI (Woo and Allen, 2014) using Mplus 8.3. We first established two profiles and then gradually increased the profiles until the model fitting index no longer improved (Nylund et al., 2007). For the model fitting index, referring to previous studies (Lo, 2001; Gabriel et al., 2015), we used the following: the log likelihood (LL), the free parameter (FP), the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the sample-size-adjusted BIC (SSA–BIC), entropy, the bootstrap likelihood ratio test (BLRT), and the Lo-Mendell-Rubin likelihood ratio test (LMR). We consider the theoretical significance of the model and model indicators to identify the best-fitting model (Foti et al., 2012). The number of retained profiles should consider both the theoretical meaning of medical AI subpopulations and model indicators [lower LL, AIC, BIC, and SSA–BIC; higher entropy; and significant LMR ($p < 0.05$)].

Study 1: Results

Identification of profiles

Table 2 provides descriptive information for the study variables. As shown in Table 3, the 3-profile solution had low LL, AIC, and SSA-BIC. In addition, the elbow plot of BIC (Figure 1) shows that the slope of the BIC curve flattens around three profiles. Moreover, the 3-profile had significant LMR, unlike other solutions that had lower LL, AIC, and SSA-BIC. More importantly, the 3-profile had theoretical meaning for medical AI. Theoretically, as the number of profiles increased,

TABLE 2 Means, standard deviations, and correlations of study 1 variables ($N = 328$).

	M (*SD*)	1	2	3	4	5	6
1. Age	29.77 (8.18)	–					
2. Sex	1.62 (0.49)	−0.21**					
3. Education years	15.61 (1.75)	−0.06	−0.01				
4. Negative attitudes	2.34 (0.71)	0.06	−0.05	0.01			
5. Objective knowledge	1.12 (0.83)	−0.08	−0.01	0.07	0.17**		
6. Subjective knowledge	3.40 (0.74)	0.06	−0.07	0.04	0.03	0.05	
7. Behavioral intentions	3.72 (0.61)	0.07	−0.07	0.03	−0.16**	0.01	0.45**

Sex (1 = male; 2 = female).
**p < 0.01.

TABLE 3 Fit statistics for profile solutions in study 1 and study 2.

Number of profiles	LL	FP	AIC	BIC	SSA-BIC	Entropy	BLRT (*p*)	LMR (*p*)
Study 1 ($N = 328$)								
2	−7,543.868	52	15,191.735	15,388.972	15,224.029	0.917	0.0000	0.0368
3	−7,328.101	70	14,796.202	15,061.713	14,839.675	0.884	0.0000	0.0214
4	−7,259.976	88	14,695.953	15,029.738	14,750.604	0.878	0.0000	0.6511
5	−7,183.407	106	14,578.814	14,980.873	14,644.644	0.855	0.0000	0.3155
6	−7,126.079	124	14,500.159	14,970.492	14,577.168	0.876	0.0000	0.2367
7	−7,081.915	142	14,447.831	14,986.439	14,536.018	0.884	0.0000	0.8302
8	−7,061.320	160	14,442.640	15,049.522	14,542.006	0.892	0.2353	0.3077
Study 2 ($N = 388$)								
2	−8,767.753	52	17,639.506	17,845.479	17,680.487	0.923	0.0000	0.0002
3	−8,573.748	70	17,287.497	17,564.767	17,342.663	0.927	0.0000	0.1050
4	−8,463.695	88	17,103.389	17,451.958	17,172.741	0.858	0.0000	0.4288
5	−8,367.262	106	16,946.523	17,366.390	17,030.061	0.858	0.0000	0.2004
6	−8,295.602	124	16,839.204	17,330.369	16,936.928	0.872	0.0000	0.5922
7	−8,229.962	142	16,743.925	17,306.387	16,855.834	0.889	0.0000	0.7110
8	−8,189.911	160	16,699.823	17,333.583	16,825.917	0.899	0.0000	0.6499

these solutions contained redundant profiles of medical AI that modeled variants of the three main profiles. Thus, the 3-profile model can ensure theoretical parsimony while also meeting the statistical criterion. Together, these theoretical, visual and statistical considerations suggest that the 3-profile model is the best model with our data.

Table 4 shows descriptive information of the retained profiles. As shown in Figure 2, among the 328 people who completed the questionnaires, 92 (28%) participants were classified into subtype 1, which had the lowest objective and subjective knowledge of medical AI, yet they also had a middle level of negative attitudes toward medical AI and the lowest level of behavioral intention regarding medical AI use.

A total of 191 (58%) participants were classified as subtype 2. The participants in this subtype showed a moderate level of objective and subjective knowledge of medical AI, yet they had the lowest negative attitudes toward medical AI and the highest behavioral intention toward medical AI use.

Forty-five (14%) participants were classified as subtype 3. The participants in this subtype showed a high level of objective and subjective knowledge of medical AI, yet they had the highest level of negative attitudes toward medical AI and a middle level of behavioral intention toward medical AI use.

Study 2: Replication and theoretical development of KAB profiles of medical AI

We first intended to replicate the main results of Study 1; thus, we expected to find the same 3 profiles of medical AI. Accordingly, we seek to explore the following question:

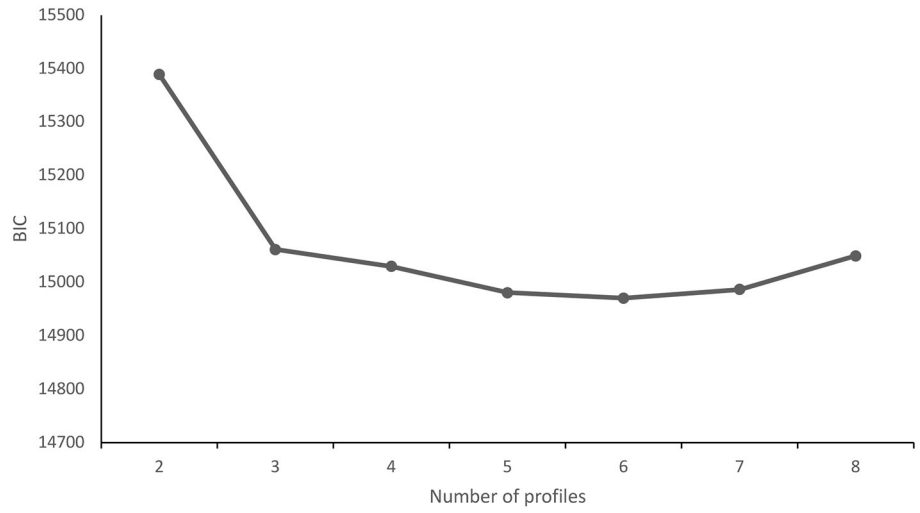

FIGURE 1
Goodness of fit of the BIC. The y-axis represents BIC (Bayesian information criterion); the x-axis represents the number of profiles (starting from 2).

Research question 1: Will three distinct KAB profiles of medical AI emerge?

We also expected to extend Study 1 by examining the antecedents of KAB profiles of medical AI in Study 2. When exploring the KAB profiles, it is critical to identify factors that can predict KAB profile membership. Previous research argues that individuals' trust and risk perception of medical AI predict their reluctance to use medical AI (Esmaeilzadeh, 2020). Thus, we pose the following question:

Research question 2: Do trust perception of medical AI and risk perception of medical AI predict KAB profile membership?

Participants and procedure

We recruited 388 participants. No participants were excluded. The participants provided informed consent and completed the survey. Table 5 provides demographic information on our sample.

Measures

Objective knowledge of medical AI

We used the same three-item multiple choice test to measure the participants' objective understanding of medical AI as in Study 1.

Subjective knowledge of medical AI

We used the same three items to measure subjective knowledge of medical AI as in Study 1 ($\alpha = 0.75$).

Negative attitudes of medical AI

We used the same 8 items to measure negative attitudes toward medical AI as in Study 1 ($\alpha = 0.85$).

Behavioral intention of medical AI use

We used the same 5 items to measure behavioral intention regarding medical AI use as in Study 1 ($\alpha = 0.79$).

Trust perception of medical AI

We measured the behavioral intention toward medical AI use using a 5-item scale (Esmaeilzadeh, 2020). The participants were asked to indicate their level of agreement with the statements (1 = "Strongly disagree," 5 = "Strongly agree"). A sample item is as follows: "I trust the medical AI algorithms used in healthcare" ($\alpha = 0.77$).

Risk perception of medical AI

We measured the behavioral intention toward medical AI use using a 5-item scale (Esmaeilzadeh, 2020). The participants were asked to indicate their level of agreement with the statements (1 = "Strongly disagree," 5 = "Strongly agree"). A sample item is as follows: "The risk of using medical AI-based tools for medical purposes is high" ($\alpha = 0.85$).

10.3389/frai.2022.1006173

TABLE 4 Descriptive information of the retained profiles in study 1 and study 2.

Profiles	% of sample	Objective knowledge	Subjective knowledge	Negative attitude	Behavior intention
Study 1 ($N = 328$)		*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
1	28%	0.97 (0.76)	2.72 (0.65)	2.45 (0.47)	3.05 (0.53)
2	58%	1.10 (0.83)	3.65 (0.59)	1.98 (0.42)	4.03 (0.36)
3	14%	1.51 (0.84)	3.73 (0.61)	3.66 (0.45)	3.80 (0.51)
Study 2 ($N = 388$)					
1	24%	1.42 (0.88)	2.99 (0.67)	2.63 (0.44)	3.13 (0.61)
2	68%	1.45 (0.76)	3.64 (0.67)	1.84 (0.38)	4.15 (0.38)
3	8%	1.58 (0.85)	3.56 (0.65)	3.76 (0.40)	3.63 (0.62)

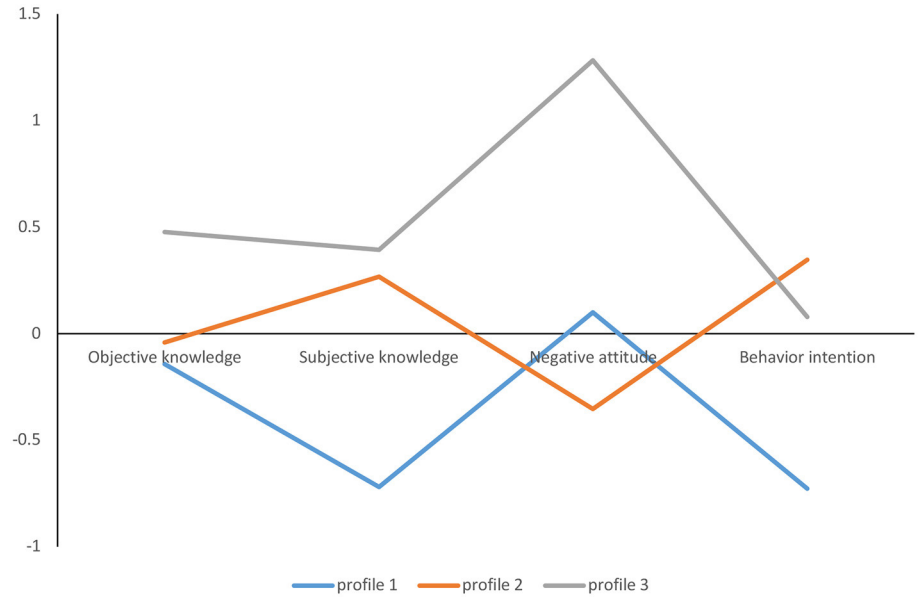

FIGURE 2
Latent profiles KAB profiles of medical AI. The y-axis refers to the mean Z score of the participants' objective knowledge, subjective knowledge, negative attitude, and behavioral intentions.

Study 2: Results

Replicating profiles

Table 6 reports descriptive information for our variables. Table 3 reports fit information for profile solutions. Table 4 illustrates descriptive information for the retained three-profile solution. The three-solution was chosen because it had lower AIC, BIC, and SSA-BIC. It also had the highest entropy. Moreover, the elbow plot of BIC (Figure 3) shows that the slope of the curve flattens around three profiles. Theoretically, when the number of profiles of medical AI increased, these profile solutions contained redundant profiles that included variants of the three main medical AI profiles. Thus, to ensure theoretical parsimony, we identified the three-profile solution as the best-fitting model for our data.

For research question 1, we replicated the three profiles as in Study 1. As shown in Figure 4, among the 388 people who completed the questionnaires, 93 (24%) participants were classified into subtype 1, which had the lowest objective and subjective knowledge of medical AI, yet they also had a middle level of negative attitudes toward medical AI and the lowest level of behavioral intention toward medical AI use.

A total of 264 (68%) participants were classified into subtype 2. The participants in this subtype showed a moderate level of objective and subjective knowledge of medical AI, yet they had the lowest negative attitudes toward medical AI and the highest behavioral intention toward medical AI use.

TABLE 5 Demographic characteristics ($N = 388$).

Variables	Frequencies (percentages)
Age	
Mean (SD)	30.48 (8.47)
Sex	
Male	158 (40.7%)
Female	230 (59.3%)
Education year	
Mean (SD)	15.82 (1.67)
Occupation	
Full-time student	47 (12.1%)
Production	14 (3.6%)
Sales	30 (7.7%)
Public relations	15 (3.9%)
Customer service	7 (1.8%)
Administration	36 (9.3%)
Human resources	12 (3.1%)
Finance	36 (9.3%)
Clerical work	21 (5.4%)
Research	68 (17.5%)
Management	40 (10.3%)
Teaching	17 (4.4%)
Consulting	0 (0%)
Professional services (e.g., journalism, law)	31 (8.0%)
Other	14 (3.6%)

Thirty-one (8%) participants were classified as subtype 3. The participants in this subtype showed a high level of objective and subjective knowledge of medical AI, yet they had the highest level of negative attitudes toward medical AI and a middle level of behavioral intention toward medical AI use.

Examination of antecedents

Regarding antecedents, following previous studies (Vermunt, 2010; Asparouhov and Muthén, 2014), we used the RESTEP to test which variables are related to the profiles of medical AI. As shown in Table 7, we found that trust perception of medical AI, risk perception of medical AI, whether AI would replace my job, AI's benefit in medicine, and whether AI cooperates with humans are significant antecedents of the KAB profile membership of medical AI. Specifically, those perceiving a higher trust perception of medical AI were more likely to be in profile 2 [odds ratios (OR) = 14.91, $p = 0.027$] than in profile 1. Those perceiving a higher risk perception of medical AI were more likely to be in profiles 1 [odds ratios (OR) = 3.22, $p = 0.026$] and 3 (OR = 4.70, $p = 0.048$) than in profile 2. Those perceiving a higher perception of medical AI replacing my job were less likely to be in profile 3 [odds ratios (OR) = 0.17, $p = 0.000$] than in profile 2. Those perceiving a higher perception of medical AI's benefit in medicine were less likely to be in profiles 1 [odds ratios (OR) = 0.41, $p = 0.000$] and 3 (OR = 0.27, $p = 0.000$) than in profile 2. Those perceiving a higher trust perception of medical AI were more likely to be in profiles 2 [odds ratios (OR) = 14.91, $p = 0.027$] and 3 (OR = 2.84, $p = 0.048$) than in profile 1. Those perceiving a higher cooperation between medical AI and humans were less likely to be in profile 1 [odds ratios (OR) = 0.44, $p = 0.00$] than in profile 3. Those perceiving a higher cooperation between medical AI and humans were less likely to be in profile 1 [odds ratios (OR) = 0.54, $p = 0.00$] than in profile 2. We found no other significant results.

Discussion

Summary of the findings

The results of this study showed that there is heterogeneity in people's medical AI use. We identified 3 profiles based on objective knowledge and subjective knowledge of and negative attitudes and behavioral intentions toward medical AI. First, the participants in profile 1 had the lowest objective and subjective knowledge of medical AI, yet they also had a middle level of negative attitudes toward medical AI and the lowest level of behavioral intention regarding medical AI use. Second, the participants in profile 2 showed a moderate level of objective and subjective knowledge of medical AI, yet they had the lowest negative attitudes toward medical AI and the highest behavioral intention toward medical AI use. Third, the participants in profile 3 showed a high level of objective and subjective knowledge of medical AI, yet they had the highest level of negative attitudes toward medical AI and a middle level of behavioral intention toward medical AI use.

Theoretical implications

Our research makes a variety of theoretical contributions. First, by taking a person-centered approach that categorized people into different profiles based upon their objective and subjective knowledge of and negative attitudes and behavioral intentions toward medical AI, our results depict a more holistic picture of people who are reluctant to use medical AI (Wang and Hanges, 2011). Most of our sampled individuals fell into profile 2, supporting the KAB model's hypothesis that knowledge, attitudes and behavior are related (Yi and Hohashi, 2018). That is, individuals with high knowledge have a low negative attitude and high behavioral intentions toward objects. The existence of profile 3 departs from the argument of the KAB model and the predominant variable-centered

TABLE 6 Means, standard deviations, and correlations of study 2 variables ($N = 388$).

	M (*SD*)	1	2	3	4	5	6
1. Age	30.48 (8.47)	–					
2. Sex	1.59 (0.49)	−0.13*					
3. Education (years)	15.82 (1.67)	−0.20**	−0.07				
4. Negative attitudes	2.18 (0.70)	−0.03	0.02	−0.01			
5. Objective knowledge	1.45 (0.79)	−0.02	0.00	−0.01	0.06		
6. Subjective knowledge	3.48 (0.72)	0.05	0.08	0.10	−0.14**	−0.02	
7. Behavioral intentions	3.86 (0.64)	0.11*	−0.05	−0.05	−0.44**	−0.02	0.40**

Sex (1 = male; 2 = female).
*p < 0.05.
**p < 0.01.

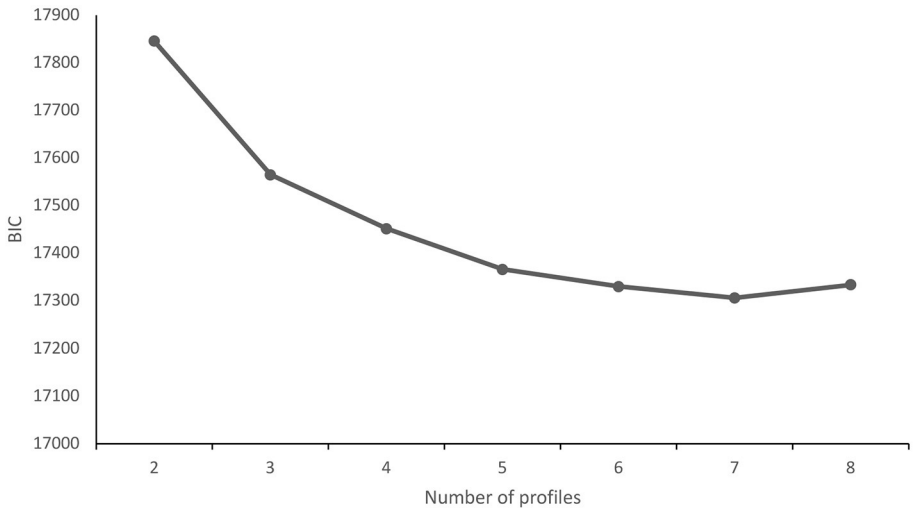

FIGURE 3
Goodness of fit of the BIC. The y-axis represents BIC (Bayesian information criterion); the x-axis represents the number of profiles (starting from 2).

method that suggests links among knowledge, attitude and behavior instead highlighting the idea that these attributes and actions separately shape individuals' holistic picture of medical AI.

Second, while the KAB model (Chaffee and Roser, 1986; Abera, 2003) provides a useful lens through which to view the diversity of people's medical AI use, our study also gives back to this theory by revealing the shortcomings of this model. Notably, across the two samples, we did not observe a profile characterized by a middle level of objective and subjective knowledge of and middle levels of negative attitudes and behavioral intention toward medical AI, which could be a reasonable prediction derived from the KAB model. One potential explanation for this pertains to the complexity and heterogeneity of medical AI use (Cadario et al., 2021). That is, the barriers to medical AI use are not a simple phenomenon that can be completely explained by the KAB model. Instead, there is considerable heterogeneity in individuals reluctant to use medical AI. Thus, when considering complex phenomena such as medical AI use, we cannot simply use KAB to apply to this context and come to a simple conclusion.

Third, our work supports and extends the KAB model on the role of trust perception and risk perception in shaping individuals' medical AI use by developing and operationalizing a coherent framework of antecedents of medical AI use profiles. Consistent with the AI literature (Esmaeilzadeh, 2020; Dwivedi et al., 2021), trust and risk perception, AI replacing the jobs of humans, AI's benefit in medicine and AI's cooperation with humans were differentially related to medical AI use profile.

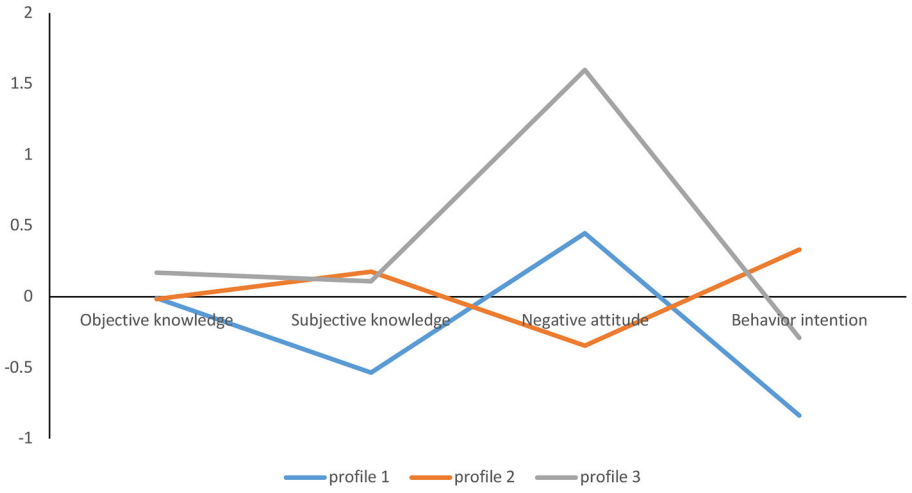

FIGURE 4
Latent profiles KAB profiles of medical AI. The y-axis refers to the mean Z score of the participants' objective knowledge, subjective knowledge, negative attitudes, and behavioral intentions.

TABLE 7 Three-step results for antecedents (RESTEP) in study 2.

Antecedents	Profile 1 v. 2	Profile 1 v. 3	Profile 2 v. 3
Trust of medical AI	−2.702***	−1.045*	1.657**
Risk of medical AI	1.169***	−0.379	−1.548***
AI replacing job	−0.444	1.330*	1.774***
AI benefit in medicine	−0.885**	0.408	1.293**
AI cooperation with humans	−0.608*	−0.820*	−0.212

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

Practical implications

Our study results provide many practical insights indicating the importance of helping individuals, media communicators, medical doctors and enterprise managers make sense of the complexity and heterogeneity of individuals' reluctance to use medical AI. For example, medical doctors could realize that some people exhibit consistent knowledge of and attitudes and behaviors toward medical AI, but others exhibit more variability in these domains, so there is no way to reach a simple and general conclusion about this subject. Importantly, our results highlight the importance of recognizing that there may be disassociation between someone's knowledge of and negative attitudes toward medical AI. Decision makers should be cautious when giving advice to individuals even if the individuals appear to have high knowledge of medical AI. Last, decision makers and policy makers may be able to create personalized intervention and dissemination programs to improve people's knowledge of AI, especially their subjective knowledge, and to help individuals in need actively adopt AI in seeking medical care in the future.

Limitations and future directions

Our research has several limitations, which may offer fruitful directions for future research. First, future research may build upon our findings to explore whether the three identified profiles of medical AI exist and new profile(s) emerge in different cultural contexts with different samples to address the representativeness of the sample. Second, as people's knowledge of and negative attitudes and behavioral intentions toward medical AI might change over time, it is possible to employ latent transition analysis (Collins and Lanza, 2009) to address the shift in the KAB profile of medical AI. Third, in our study, objective knowledge and subjective knowledge were consistent, and there was no significant difference in shaping the profile of medical AI. This may be because our sample is the general public, and there is no significant difference between their objective and subjective knowledge of medical AI. However, for professionals, such as doctors, it is still worth exploring the effects of age in shaping people's reluctance to use medical AI in depth.

Conclusion

The burgeoning AI literature has been limited in its understanding of the diversity in people's reluctance

to use medical AI. We used LPA to better understand the heterogeneity of people's reluctance to use medical AI regarding their knowledge, negative attitudes and behavioral intentions. Our results demonstrated that different medical AI profiles consistently exist, and it is helpful to use a person-centered approach to better understand the complexity of obstacles in people's reluctance to use medical AI.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethical Committee of Jinan University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

HW, KL, and LH conceived and designed the research. HW and LH performed the research. HW, KL, QS, and LG analyzed the data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Program of National Natural Science Foundation of China (Grant Numbers 72174075 and 71801109) and Humanity and Social Science Youth Foundation of Ministry of Education of China (Grant Numbers 19YJCZH073).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.1006173/full#supplementary-material

References

Abdullah, R., and Fakieh, B. (2020). Health care employees' perceptions of the use of artificial intelligence applications: survey study. *J. Med. Intern. Res.* 22, e17620. doi: 10.2196/17620

Abera, Z. (2003). Knowledge, attitude and behavior (KAB) on HIV/AIDS/STDs among workers in the informal sector in Addis Ababa. *Ethiop. J. Health Dev.* 17, 53–61. doi: 10.4314/ejhd.v17i1.9781

Asparouhov, T., and Muthén, B. (2014). Auxiliary variables in mixture modeling: three-step approaches using M plus. *Struct. Eq. Model. Multidiscip. J.* 21, 329–341. doi: 10.1080/10705511.2014.915181

Cadario, R., Longoni, C., and Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nat. Hum. Behav.* 5, 1636–1642. doi: 10.1038/s41562-021-01146-0

Chaffee, S. H., and Roser, C. (1986). Involvement and the consistency of knowledge, attitudes, and behaviors. *Commun. Res.* 13, 373–399. doi: 10.1177/009365086013003006

Collins, L. M., and Lanza, S. T. (2009). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, NJ: John Wiley & Sons.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., et al. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int. J. Inform. Manag.* 57, 101994. doi: 10.1016/j.ijinfomgt.2019.08.002

Eastwood, J., Snook, B., and Luther, K. (2012). What people want from their professionals: attitudes toward decision-making strategies. *J. Behav. Decis. Mak.* 25, 458–468. doi: 10.1002/bdm.741

Ekehammar, B., and Akrami, N. (2003). The relation between personality and prejudice: a variable- and a person-centred approach. *Eur. J. Pers.* 17, 449–464. doi: 10.1002/per.494

Esmaeilzadeh, P. (2020). Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC Med. Inform. Decis. Mak.* 20, 170. doi: 10.1186/s12911-020-01191-1

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi: 10.1038/nature21056

Foti, R. J., Bray, B. C., Thompson, N. J., and Allgood, S. F. (2012). Know thy self, know thy leader: contributions of a pattern-oriented approach to examining leader perceptions. *Leadership Quart.* 23, 702–717. doi: 10.1016/j.leaqua.2012.03.007

Gabriel, A. S., Daniels, M. A., Diefendorff, J. M., and Greguras, G. J. (2015). Emotional labor actors: a latent profile analysis of emotional labor strategies. *J. Appl. Psychol.* 100, 863–879. doi: 10.1037/a0037408

Hao, K. (2020). *Doctors are Using AI to Triage COVID-19 Patients. The Tools may be Here to Stay*. Boston, MA: MIT Technology Review.

Hohashi, N., and Honda, J. (2015). Concept development and implementation of family care/caring theory in concentric sphere family environment theory. *Open J. Nurs.* 5, 749–757. doi: 10.4236/ojn.2015.59078

Hollander, J. E., and Carr, B. G. (2020). Virtually perfect? Telemedicine for covid-19. *N. Engl. J. Med.* 382, 1679–1681. doi: 10.1056/NEJMp2003539

Kemm, J. R., and Close, A. (1995). *Health Promotion: Theory and Practice*. London: Macmillan International Higher Education.

Lo, Y. (2001). Testing the number of components in a normal mixture. *Biometrika* 88, 767–778. doi: 10.1093/biomet/88.3.767

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *J. Cons. Res.* 46, 629–650. doi: 10.1093/jcr/ucz013

Marsh, H. W., Lüdtke, O., Trautwein, U., and Morin, A. J. S. (2009). Classical latent profile analysis of academic self-concept dimensions: synergy of person- and variable-centered approaches to theoretical models of self-concept. *Struct. Eq. Model. Multidiscip. J.* 16, 191–225. doi: 10.1080/10705510902751010

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Eq. Model. Multidiscip. J.* 14, 535–569. doi: 10.1080/10705510701575396

Price, W. N. (2018). Big data and black-box medical algorithms. *Sci. Transl. Med.* 10, eaao5333. doi: 10.1126/scitranslmed.aao5333

Promberger, M., and Baron, J. (2006). Do patients trust computers? *J. Behav. Decis. Mak.* 19, 455–468. doi: 10.1002/bdm.542

Schepman, A., and Rodway, P. (2020). Initial validation of the general attitudes towards artificial intelligence scale. *Comput. Hum. Behav. Rep.* 1, 100014. doi: 10.1016/j.chbr.2020.100014

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Polit. Anal.* 18, 450–469. doi: 10.1093/pan/mpq025

Wang, M., and Hanges, P. J. (2011). Latent class procedures: applications to organizational research. *Org. Res. Methods* 14, 24–31. doi: 10.1177/1094428110383988

Woo, S. E., and Allen, D. G. (2014). Toward an inductive theory of stayers and seekers in the organization. *J. Bus. Psychol.* 29, 683–703. doi: 10.1007/s10869-013-9303-z

Woo, S. E., Jebb, A. T., Tay, L., and Parrigon, S. (2018). Putting the "person" in the center. *Org. Res. Methods* 21, 814–845. doi: 10.1177/1094428117752467

Wosik, J., Fudim, M., Cameron, B., Gellad, Z. F., Cho, A., Phinney, D., et al. (2020). Telehealth transformation: COVID-19 and the rise of virtual care. *J. Am. Med. Inform. Assoc.* 27, 957–962. doi: 10.1093/jamia/ocaa067

Xu, L., and Yu, F. (2019). Factors that influence robot acceptance. *Chin. Sci. Bull.* 65, 496–510. doi: 10.1360/TB-2019-0136

Yi, Q., and Hohashi, N. (2018). Comparison of perceptions of domestic elder abuse among healthcare workers based on the knowledge-attitude-behavior (KAB) model. *PLoS ONE* 13, e0206640. doi: 10.1371/journal.pone.0206640

Zyphur, M. J. (2009). When mindsets collide: switching analytical mindsets to advance organization science. *Acad. Manag. Rev.* 34, 677–688. doi: 10.5465/AMR.2009.44885862

frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 03 March 2023
DOI 10.3389/frai.2023.1148227

OPEN ACCESS

EDITED BY
Christos Troussas,
University of West Attica, Greece

REVIEWED BY
Zhi Liu,
Central China Normal University, China

*CORRESPONDENCE
Mitchell J. Nathan
✉ mnathan@wisc.edu

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 19 January 2023
ACCEPTED 14 February 2023
PUBLISHED 03 March 2023

CITATION
Nathan MJ (2023) Disembodied AI and the limits to machine understanding of students' embodied interactions. *Front. Artif. Intell.* 6:1148227. doi: 10.3389/frai.2023.1148227

Disembodied AI and the limits to machine understanding of students' embodied interactions

Mitchell J. Nathan*

MAGIC Lab, Wisconsin Center for Education Research, Educational Psychology Department, School of Education at the University of Wisconsin–Madison, Madison, WI, United States

The *embodiment turn* in the Learning Sciences has fueled growth of multimodal learning analytics to understand embodied interactions and make consequential educational decisions about students more rapidly, more accurately, and more personalized than ever before. Managing demands of complexity and speed is leading to growing reliance by education systems on disembodied artificial intelligence (dAI) programs, which, ironically, are inherently incapable of interpreting students' embodied interactions. This is fueling a potential *crisis of complexity*. *Augmented intelligence* systems offer promising avenues for managing this crisis by integrating the strengths of omnipresent dAI to detect complex patterns of student behavior from multimodal datastreams, with the strengths of humans to meaningfully interpret embodied interactions in service of consequential decision making to achieve a balance between complexity, interpretability, and accountability for allocating education resources to children.

KEYWORDS

artificial intelligence, augmented intelligence, cognitive science, embodied learning, foundation models, learning sciences, multimodality

1. Introduction

The primary objective of this *Perspectives* article is to expose a looming crisis of complexity: educational systems are becoming more dependent on artificial intelligence (AI) programs to make consequential decisions about learning and learners from rich streams of multimodal data that emerge from many sources, including students' embodied interactions. However, disembodied AI (dAI) programs–I argue–are fundamentally incapable of understanding people's embodied interactions in the ways that humans understand them. Furthermore, the emergent dAI models are of such complexity that end users (and often the original programmers) cannot understand the models or recreate the chain of reasoning that led to these decisions. Therefore, dAIs should not be directing consequential educational decisions affecting the lives of children. The secondary objective is to offer potential paths forward from this crisis. One promising approach is the development of "augmented intelligence" systems (AISs) that amplify human performance using dAI resources while relying ultimately on human decision making.

2. Theoretical framework: The embodied turn and growth of multimodal learning analytics

2.1. The embodied turn in the learning sciences and education

Empirical evidence and arguments from philosophy, psychology, neuroscience, education, and critical theorists in education effectively dismantle the view of learning as information processing of ungrounded symbol systems by dAI that are amodal (i.e., non-sensorial), arbitrary (i.e., non-historical and non-cultural), and abstract (i.e., ungrounded) (Harnad, 1990; Varela et al., 1991; Glenberg, 1997; Shapiro, 2019). To the contrary, humans make meaning of events, ideas, and cultural and scientific inscriptions by grounding them to their sensorimotor experiences that are interpreted within sociocultural and historical contexts (Wilson, 2002; Barsalou, 2008; Newen et al., 2018).

In psychology, Glenberg and Robertson (2000) found that human readers judge the sensibility of sentences based on the sensorimotor affordances invoked by the actions described in the sentences, rather than their lexical interconnections in high-dimensional spaces, as modeled by dAI systems widely applied in education areas such as automated essay grading (LSA; Burgess and Lund, 1997; Landauer and Dumais, 1997).

Neural imaging data show that reading words with motor associations—such as kick, lick, and pick—selectively activates the motor areas of the brain for one's feet, tongue, and fingers, respectively (Pulvermüller, 2005). Botox patients whose injections temporarily paralyze the facial corrugator supercilli muscle used in frowning showed selective impairment in processing sentences that invoke anger but not those that invoked joy or were emotionally neutral (Havas et al., 2010).

Critical theorists in education reject the disembodied view that neglects the central role of culture in language, thinking, symbols, and emotion for educational attainment. McKinney de Royston et al. (2020) expressly identify the essential nature of embodied cultural experiences by framing learning as rooted in bodies and brains that are embedded in social and cultural practices and shaped by lifelong culturally organized activities.

Drawing on these critiques, some education scholars conclude that the knowledge and educational practices of students and teachers are fundamentally determined by people's individual and collective embodied processes in order to make sense of their school-based learning experiences (e.g., Shapiro and Stolz, 2019; Nathan, 2021; Macrine and Fugate, 2022). This has led to innovative designs in embodied learning through educational technology (Papert, 2020; Abrahamson and Lindgren, 2022), embodiment in AI and education (Timms, 2016) and embodied conversational agents (Cassell, 2001) that promote student learning and intellectual development.

2.2. Growth of multimodal learning analytics

With the embodiment turn has emerged methods for collecting and analyzing *multimodal data* to model embodied interactions (Worsley and Blikstein, 2018; Abrahamson et al., 2021). These include data for analyzing gestures (Closser et al., 2021), eye gaze (Schneider and Pea, 2013; Shvarts and Abrahamson, 2019), facial expression (Monkaresi et al., 2016; Sinha, 2021), grip intensity (Laukkonen et al., 2021), and so on, coupled with traditional statistical methods, qualitative methods, and deep learning algorithms that model human behavior based on massive amounts of mouse click and text-based data (e.g., Facebook's DeepText, Google's RankBrain). This shift in research methods has been enabled by the proliferation of low-cost, high-bandwidth cameras and sensors that track biometrics, facial, and body movement that supplement field notes, speech, text chat, and click log data (Schneider and Radu, 2022).

Work with multimodal data has historically been labor-intensive and subject to the severely limited processing capacities of humans that constrain the amount of data under consideration, its dimensionality, and the cycle time between data collection, interpretation, and action. This restricted the ability to use multimodal data to identify latent patterns and inform practitioners in real time about embodied interactions relevant to on-task and off-task behavior. Some of the forces that propelled educational data mining and learning analytics (Aldowah et al., 2019; Baker and Siemens, 2022) have motivated the creation of more efficient data analytic tools and algorithms to process massive multimodal corpora (e.g., An et al., 2019; Järvelä et al., 2019). This is leading to the emergence of new methodological practices of *multimodal learning analytics and data mining* (hereafter MMLA; Blikstein and Worsley, 2016).

3. Analytic method and evidence: The disconnect between dAI and human meaning making

An analysis of the computational architectures of classical and contemporary AI systems that underly the tools for MMLA reveals that they are fundamentally incapable of understanding the meaning of people's embodied interactions, even as they give the appearance of mimicking intelligent embodied behavior.

Classical, symbol-based AI systems were designed and implemented by human programmers to emulate human intelligence. The arbitrary, amodal, and abstract nature of these symbol systems was a feature, not a bug, and key to the power of these computational algorithms to operate consistently and efficiently, across a wide range of domains. For example, semantic nets presumably could model any organization of memory (Collins and Loftus, 1975). Although classical AI systems excelled at the analytic tasks that are the signature of adult intellect, such as complicated calculations and hierarchical inference-making, they were wholly inadequate at performing culturally familiar tasks well within reach of children, such as balance, face recognition, and

basic social interactions (e.g., Resnick, 1987) and struggled to be adaptive in the face of task, environmental, and user variation.

Connectionist architectures arose that addressed many limitations of classical AI. Often, these drew on parallel and distributed forms of computation that adapted to training experiences through the adjustment of strengths of connections among simple nodes in large networks, mediated by hidden layers (McClelland et al., 1986; Rumelhart et al., 1988). These systems excelled at simple pattern learning and prediction, and at many of the sensorimotor skills that eluded early symbolic AI systems. Yet these connectionist systems found many symbol analytic tasks cumbersome. These systems depended heavily on carefully cultivated training sets and pre-coded sensory inputs for successful learning, underscoring their disembodied nature.

New approaches arose that exploited high-dimensional spaces for computing variability and similarity, greatly expanding the training sets they could accommodate and the complexity of the associations they could encode (e.g., Burgess and Lund, 1997; Landauer and Dumais, 1997). Thus, attention in AI development turned to the importance of training experiences and the sheer number of nodes and inter-nodal connections used by these systems.

This fueled the current movement to Foundation AI systems such as BERT, GPT-3, and DALL-E that are built to accommodate enormous training corpora with massive numbers of internodal connections (Bommasani et al., 2021). Foundation AI systems are designed to learn on their own and be adaptive to completely new, untrained conditions—often in ways that their creators cannot foresee. For example, GPT-3 is built on 175 billion parameters trained on 570 Gigabytes of text. GPT-3 can learn to write original essays, produce computer code, and generate reasonable responses to novel discourse (not just novel syntactic structures) it has never been trained on.

Still, these systems are working from disembodied patterns extracted from the regularities of how words and images occur in the training datastreams. GPT-3, as a representative example, "lacks intentions, goals, and the ability to understand cause and effect" [Percy Liang, Director of Stanford's Center for Research on Foundation Models (CRFM), in CRFM, 2021] that naturally come from human being's embodied interactions with one's environment and other people. Newer language models, such as ChatGPT, are based on GPT-3 architecture and develop their language generation and comprehension capabilities through these same basic analytic methods, coupled with a mechanism of Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022) from human labelers. Despite its fascination in the media, RLHF has significant limitations as noted by the developers (Ouyang et al., 2022). Its future performance is based on a number of subjective and untested sources of human bias; specifically: unaccounted for biases of the human labelers and the researchers who initially developed the instructions used by the labelers; the prompts provided by the developers and early users; and that the same human biases are present in the training and model evaluation process. Furthermore, foundation models like GPT-3, ChatGPT, and the like are completely opaque: the creators do not know how the models will work in new domains and cannot predict the future interactions of their creations What's more, in what is both a profound strength and a serious weakness, architectural and training decisions made early on influence a system throughout its lifetime. Thus, when key considerations such as embodiment are neglected, one cannot simply go back and retrofit changes (Bommasani et al., 2021).

These issues of disembodiment, opaqueness, and developmental fixedness all converge to shape a distorted image of what the educational community should be drawn to. As Liang notes in a recent webinar (CRFM, 2021), ideally, "the ethical and social awareness needs to be integrated into the technological development." However, the norm for social and ethical considerations is to follow *after* the technology is built, trained, and deployed. Liang laments "At that point I think it's too late [Because of emergence and homogenization] some of the critical decisions have been made already, in a structural way" (CRFM, 2021).

Despite their enormous computing power, dAI programs for MMLA are fundamentally incapable of deriving human-centered meaning from embodied interactions. dAI programs fail along philosophical grounds to achieve intentionality (Searle, 1980). Instead, they generate ungrounded models of behavior linked to high-dimensional statistical regularities of behavior, rather than the meaningful embodied experiences they purport to model (Harnad, 1990). They fall short phenomenologically by relying on mathematical redescriptions that intervene between sensation and action (Gallagher, 2018). And the symbol structures they generate to describe human behavior have no cultural or historical bases (McKinney de Royston et al., 2020). As Barsalou (1999, p. 608) states, "computers should not be capable of implementing a human conceptual system, because they do not have the requisite sensory-motor systems for representing human concepts."

4. Urgency of the problem of dAI in educational decision making

A variety of automated detectors have been developed that use non-invasive methods to classify students' emotional states, engagement, and cognitive presence during their participation in on-line classes (e.g., Baker et al., 2010; Liu et al., 2019, 2023).

The increasing availability of multimodal data has coincided with growing expectations for computers to deliver data-driven, real-time directives for education, such as personalized learning (Walkington, 2013) and assessment, added pressures from a global pandemic that disrupted standard, in-person learning, and a lack of oversight or regulation on the access and use of such data by machines in educational settings (Crawford, 2021). The response has been a proliferation of dAI-based solutions to traditional educational problems such as formative and summative assessment and differentiated curricula using tools, such as *4 Little Trees*, that uses eye gaze, facial expression, and body movement to make educational decisions and evaluations about student attentiveness and level of engagement (Chan, 2021; Harper et al., 2022); and systems such as TalkMoves, that collect recordings of classroom discourse but ignore students' non-verbal interactions (Suresh et al., 2021).

The urgency is that school leaders and classroom teachers looking to manage their workloads with limited resources see dAI-based systems as ready-made solutions (e.g., Tyson, 2020).

However, school leaders and teachers may be ill-informed about the actual inner workings of dAI systems and the inherent limitations of these systems to understanding people's embodied interactions in the ways that humans understand them, as described in section 2. This needs to change before educational practices become too dependent on dAI systems without proper considerations of ways to address these limitations (as outlined in the next section).

The potential risks are that students' embodied ways of expressing their reasoning are disregarded, thus providing impoverished accounts of their engagement and learning; or, that these non-verbal behaviors are incorrectly classified due to the limitations and biases built into the dAI systems. In both scenarios, dAI systems would be given authority over consequential decisions about students' educational experiences that can have lifelong consequences without adequate oversight by educators.

5. Pathways forward

Given dAI limitations, alternatives are needed to manage the complexities of embodied interactions while still offering time-sensitive, human-centered interpretations and accountable decision-making. The emergence of augmented intelligence systems (AISs; Dubova et al., 2022) in areas such as healthcare with high-levels of personal interactions (Crigger et al., 2022) and need for trust ([HLEG-AI] High-Level Expert Group on Artificial Intelligence, 2019) offer promising avenues for education. One exemplar is *detector-driven interviewing* (DDI) methods. DDIs use dAIs to continually monitor human behavior using non-invasive methods for cognitive and affective patterns that signal learning and engagement events of importance to educators (e.g., frustration detectors), then alert human researchers and practitioners of these events to trigger personalized attention, natural human interactions, and customized pedagogical support (Baker et al., 2021; Ocumpaugh et al., 2021; Hutt et al., 2022). Successful DDIs in the learning system *Betty's Brain* (Leelawong and Biswas, 2008) demonstrates its ability to improve educational responsiveness that enhances student engagement and contributes to scientific models of the cognitive and affective processes that shape learning.

6. Discussion

The *embodiment turn* in the Learning Sciences dismantles accounts of intellectual behavior that equates cognition with disembodied computation. The rise of MMLA applied to student education is fueling a quiet movement to accede human educational decision making to dAI systems. This essay uses an embodiment framework to argue that autonomous dAI systems are fundamentally incapable of understanding embodied interactions the ways that humans understand embodied interactions due to their disconnect from sensorimotor and sociocultural interactions with their environments, and therefore should not be directing consequential educational decisions. Thus, there is a looming crisis of complexity as dAI systems fundamentally incapable of understanding embodied interactions will be enlisted to manage the enormous complexities of the multimodal models used to describe those embodied interactions and make consequential educational decisions for students. Ethical and embodied AI systems seem a long way off. The time is ripe to invest in alternatives such as augmented intelligence systems that cultivate the omnipresence and computational power of dAIs with the embodied meaning making of human interpreters and decision makers (as illustrated by approaches such as detector-driven interviewing) as a means to achieve an appropriate balance between complexity, interpretability, and accountability for allocating education resources to our children.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

I wish to acknowledge the valuable comments and discussions about the ideas presented here with Ryan Baker, Stephen Hutt, and Michael Swart. Some of the ideas here were presented in my talk on 29 September, 2022 to the Augmented Intelligence (AugInt) Workshop hosted by Robert Goldstone, Mirta Galesic, Gautam Biswas, and Marina Dubova.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abrahamson, D., and Lindgren, R. (2022). "Embodiment and embodied design," in *The Cambridge Handbook of the Learning Sciences*, ed R. K. Sawyer, 3rd ed. (Cambridge: Cambridge University Press), 301–320. doi: 10.1017/9781108888295.019

Abrahamson, D., Worsley, M., Pardos, Z. A., and Ou, L. (2021). Learning analytics of embodied design: enhancing synergy. *Int. J. Child Comput. Interact.* 32:100409. doi: 10.1016/j.ijcci.2021.100409

Aldowah, H., Al-Samarraie, H., and Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: a review and synthesis. *Telemat. Inform.* 37, 13–49. doi: 10.1016/j.tele.2019.01.007

An, P., Bakker, S., Ordanovski, S., Taconis, R., Paffen, C. L., and Eggen, B. (2019). "Unobtrusively enhancing reflection-in-action of teachers through spatially distributed ambient information," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow), 1–14. doi: 10.1145/3290605.3300321

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., and Graesser, A. C. (2010). Better to be frustrated than bored: the incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *Int. J. Hum. Comput. Stud.* 68, 223–241. doi: 10.1016/j.ijhcs.2009.12.003

Baker, R. S., Nasiar, N., Ocumpaugh, J.L., Hutt, S., Andres, J.M.A.L., Slater, S., et al. (2021). "Affect-targeted interviews for understanding Student frustration," in *Proceedings of the International Conference on Artificial Intelligence and Education*, eds I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova. 52–63. doi: 10.1007/978-3-030-78292-4_5

Baker, R. S., and Siemens, G. (2022). "Learning analytics and educational data mining," in *Cambridge Handbook of the Learning Sciences*, ed R. K. Sawyer, 3rd ed. (Cambridge, UK: Cambridge University Press), 259–278. doi: 10.1017/9781108888295.016

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660. doi: 10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Blikstein, P., and Worsley, M. (2016). Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. *J. Learn. Anal.* 3, 220–238. doi: 10.18608/jla.2016.32.11

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv*. [preprint]. doi: 10.48550/arXiv.2108.07258

Burgess, C., and Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Lang. Cogn. Process.* 12, 1–34.

Cassell, J. (2001). Embodied conversational agents: representation and intelligence in user interfaces. *AI Mag.* 22, 67–83. doi: 10.1609/aimag.v22i4.1593

Chan, M. (2021). This AI reads children's emotions as they learn. *CNN Business*. February 17, 2021.

Closser, A. H., Erickson, J. A., Smith, H., Varatharaj, A., and Botelho, A. F. (2021). Blending learning analytics and embodied design to model students' comprehension of measurement using their actions, speech, and gestures. *Int. J. Child Comput. Interact.* 32:100391. doi: 10.1016/j.ijcci.2021.100391

Collins, A. M., and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychol. Rev.* 82, 407. doi: 10.1037/0033-295X.82.6.407

Crawford, K. (2021). Time to regulate AI that interprets human emotions. *Nature* 592, 167. doi: 10.1038/d41586-021-00868-5

CRFM (2021). *Workshop on Foundation Models*. Stanford University Human-Centered Artificial Intelligence. Available online at: https://crfm.stanford.edu/workshop.html

Crigger, E., Reinbold, K., Hanson, C., Kao, A., Blake, K., and Irons, M. (2022). Trustworthy augmented intelligence in health care. *J. Med. Syst.* 46, 1–11. doi: 10.1007/s10916-021-01790-z

Dubova, M., Galesic, M., and Goldstone, R. L. (2022). Cognitive science of augmented intelligence. *Cogn. Sci.* 46:e13229.

Gallagher, S. (2018). *A Well-Trodden Path: From Phenomenology to Enactivism*. Oslo: Shaun Gallagher Filosofisk Suplement.

Glenberg, A. M. (1997). What memory is for: creating meaning in the service of action. *Behav. Brain Sci.* 20, 41–50. doi: 10.1017/S0140525X97470012

Glenberg, A. M., and Robertson, D. A. (2000). Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *J. Mem. Lang.* 43, 379–401. doi: 10.1006/jmla.2000.2714

Harnad, S. (1990). The symbol grounding problem. *Phys. D Nonlinear Phenom.* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6

Harper, D. J., Ellis, D., and Tucker, I. (2022). "Covert aspects of surveillance and the ethical issues they raise," in *Ethical Issues in Covert, Security and Surveillance Research Advances in Research Ethics and Integrity*, eds R. Iphofen, and D. O'Mathúna (Bingley: Emerald Publishing Limited), 177–197. doi: 10.1108/S2398-601820210000008013

Havas, D. A., Glenberg, A. M., Gutowski, K. A., Lucarelli, M. J., and Davidson, R. J. (2010). Cosmetic use of botulinum toxin-A affects processing of emotional language. *Psychol. Sci.* 21, 895–900. doi: 10.1177/0956797610374742

[HLEG-AI] High-Level Expert Group on Artificial Intelligence (2019). *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission.

Hutt, S., Baker, R. S., Ocumpaugh, J., Munshi, A., Andres, J. M. A. L., Karumbaiah, S., et al. (2022). "Quick red fox: an app supporting a new paradigm in qualitative research on AIED for STEM," in *Artificial Intelligence in STEM Education: The Paradigmatic Shifts in Research, Education, and Technology*, eds F. Ouyang, P., Jiao, B. M. McLaren, and A. H. Alavi (Boca Raton, FL: CRC Press), 319–332. doi: 10.1201/9781003181187-26

Järvelä, S., Järvenoja, H., and Malmberg, J. (2019). Capturing the dynamic and cyclical nature of regulation: methodological progress in understanding socially shared regulation in learning. *Int. J. Comput. Support. Collab. Learn.* 14, 425–441. doi: 10.1007/s11412-019-09313-2

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211. doi: 10.1037/0033-295X.104.2.211

Laukkonen, R. E., Ingledew, D. J., Grimmer, H. J., Schooler, J. W., and Tangen, J. M. (2021). Getting a grip on insight: real-time and embodied Aha experiences predict correct solutions. *Cogn. Emot.* 35, 918–935. doi: 10.1080/02699931.2021.1908230

Leelawong, K., and Biswas, G. (2008). Designing learning by teaching agents: the Betty's Brain system. *Int. J. Artif. Intell. Educ.* 18, 181–208.

Liu, Z., Kong, X., Chen, H., Liu, S., and Yang, Z. (2023). MOOC-BERT: automatically identifying learner cognitive presence from MOOC discussion data. *IEEE Transact. Learn. Technol.* 31:1–14. doi: 10.1109/TLT.2023.3240715

Liu, Z., Yang, C., Rüdian S., Liu, S., Zhao, L., and Wang, T. (2019). Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums. *Interact. Learn. Environ.* 27, 598–627. doi: 10.1080/10494820.2019.1610449

Macrine, S., and Fugate, J. (2022). *Movement Matters: How Embodied Cognition Informs Teaching and Learning*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/13593.001.0001

McClelland, J. L., Rumelhart, D. E., and PDP Research Group (1986). *Parallel Distributed Processing* (Vol. 2). Cambridge, MA: MIT press.

McKinney de Royston, M., Lee, C., Nasir, N. S., and Pea, R. (2020). Rethinking schools, rethinking learning. *Phi Delta Kappan* 102, 8–13. doi: 10.1177/0031721720970693

Monkaresi, H., Bosch, N., Calvo, R. A., and D'Mello, S. K. (2016). Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transact. Affect. Comput.* 8, 15–28. doi: 10.1109/TAFFC.2016.2515084

Nathan, M. J. (2021). *Foundations of Embodied Learning: A Paradigm for Education*. London: Routledge. doi: 10.4324/9780429329098

Newen, A., De Bruin, L., and Gallagher, S. (Eds) (2018). *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780198735410.001.0001

Ocumpaugh, J., Hutt, S., Andres, J. M. A. L., Baker, R. S., Biswas, G., Bosch, N., et al. (2021). "Using qualitative data from targeted interviews to inform rapid AIED development," in *Proceedings of the 29th International Conference on Computers in Education* (Bangkok).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *arXiv*. [preprint] doi: 10.48550/arXiv.2203.02155

Papert, S. A. (2020). *Mindstorms: Children, Computers, and Powerful Ideas*. New York, NY: Basic books.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nat. Rev. Neurosci.* 6, 576–582. doi: 10.1038/nrn1706

Resnick, L. B. (1987). The 1987 presidential address learning in school and out. *Educ. Res.* 16, 13–54. doi: 10.3102/0013189X016009013

Rumelhart, D. E., McClelland, J. L., and PDP Research Group (1988). *Parallel Distributed Processing*, Vol. 1. Cambridge, MA: MIT press.

Schneider, B., and Pea, R. (2013). "Using eye-tracking technology to support visual coordination in collaborative problem-solving groups," in *To See the World and a Grain of Sand: Learning across Levels of Space, Time, and Scale: CSCL 2013 Conference Proceedings Volume 1—Full Papers and Symposia*, eds N. Rummel, M. Kapur, M. Nathan, and S. Puntambekar (Madison, WI: International Society of the Learning Sciences), 406–413.

Schneider, B., and Radu, I. (2022). "Augmented reality in the learning sciences," in *The Cambridge Handbook of the Learning Sciences*, ed R. K. Sawyer, 3rd ed. (Cambridge: Cambridge University Press), 340–361. doi: 10.1017/9781108888295.021

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424. doi: 10.1017/S0140525X00005756

Shapiro, L. (2019). *Embodied Cognition,* 2nd ed. New York, NY: Routledge. doi: 10.4324/9781315180380

Shapiro, L., and Stolz, S. A. (2019). Embodied cognition and its significance for education. *Theory Res. Educ.* 17, 19–39. doi: 10.1177/1477878518822149

Shvarts, A., and Abrahamson, D. (2019). Dual-eye-tracking Vygotsky: a microgenetic account of a teaching/learning collaboration in an embodied interaction technological tutorial for mathematics. *Learn. Cult. Soc. Interact.* 22, 100316. doi: 10.1016/j.lcsi.2019.05.003

Sinha, T. (2021). Enriching problem-solving followed by instruction with explanatory accounts of emotions. *J. Learn. Sci.* 31, 151–198. doi: 10.1080/10508406.2021.1964506

Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., et al. (2021). Using transformers to provide teachers with personalized feedback on their classroom discourse: the TalkMoves application. *arXiv*. [preprint]. doi: 10.48550/arXiv.2105.07949

Timms, M. J. (2016). Letting artificial intelligence in education out of the box: educational cobots and smart classrooms. *Int. J. Artif. Intell. Educ.* 26, 701–712. doi: 10.1007/s40593-016-0095-y

Tyson, M. (2020). *Educational Leadership in the Age of Artificial Intelligence* [Dissertation]. Atlanta, GA: Georgia State University.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/6730.001.0001

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: the impact of relevant contexts on performance and learning outcomes. *J. Educ. Psychol.* 105, 932. doi: 10.1037/a0031882

Wilson, M. (2002). Six views of embodied cognition. *Psychon. Bull. Rev.* 9, 625–636. doi: 10.3758/BF03196322

Worsley, M., and Blikstein, P. (2018). A multimodal analysis of making. *Int. J. Artif. Intell. Educ.* 28, 385–419. doi: 10.1007/s40593-017-0160-1

Frontiers in
Artificial Intelligence

Explores the disruptive technological revolution of AI

A nexus for research in core and applied AI areas, this journal focuses on the enormous expansion of AI into aspects of modern life such as finance, law, medicine, agriculture, and human learning.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact